딥러닝을 이용한 배달 음식점 리뷰 자동 생성

김나경[O], 조혜진[**], 이혜진[*], 정유철[*]
[O]금오공과대학교 컴퓨터공학과,
[*]금오공과대학교 컴퓨터공학과,
[**]금오공과대학교 응용수학과
e-mail: mym9926@naver.com, hjin_77@naver.com, hyeeo0o7@naver.com, jyc@kumoh.ac.kr

Automatic Review Generation for Delivery Restaurant using Deep Learning Models

Nagyeong Kim[O], Hyejin Jo[**], Hyejin Lee[*], Yuchul Jung[*]
[O]Dept. of Computer Engineering, Kumoh National Institute of Technology,
[*]Dept. of Computer Engineering, Kumoh National Institute of Technology,
[**]Dept. of Applied Mathematics, Kumoh National Institute of Technology

● 요 약 ●

본 논문에서는 딥러닝 모델 중 Keras 기반 LSTM 모델과 KoGPT-2 모델을 이용하여 학습한 결과를 바탕으로 카테고리 별 키워드 기반의 배달 음식점 리뷰를 생성하는 방법을 제안한다. 데이터는 주로 맛, 양, 배달, 가격으로 구성되어 있으며 이를 카테고리 별로 구분하였다. 또한 새롭게 생성된 텍스트는 의미와 문맥을 판단하여 기존 리뷰 데이터와 비슷하게 구현하였다. 모델마다 성능을 비교하기 위해 정량적, 정성적 평가를 진행하였다.

키워드: 딥러닝(DeepLearning), 텍스트생성(text-generation), LSTM, KoGPT-2

I. Introduction

본 논문은 카테고리 별 키워드 기반 리뷰 생성을 목적으로 한다. 음식점 리뷰는 맛, 양, 배달, 가격 등의 내용으로 이루어져 있다. 학습 데이터는 카테고리 별로 리뷰를 선별하여 데이터를 정제하고 학습시켰다. 인공지능 신경망 모델 중 하나인 Keras 기반 Seq2Seq LSTM 모델과 KoGPT-2 모델을 적용하였다. 학습시킨 모델을 토대로 카테고리 별 관련 키워드 입력 시 문장이 자동 생성되도록 하였으며, 수동 검증을 통해 문장들이 기존 리뷰와 비슷하게 생성되는지를 관찰하였다.

II. Preliminaries

1. Related works

1.1 LSTM

Long Short Term Memory networks(LSTM)[1]은 RNN의 vanishing gradient problem을 극복하기 위하여 고안되었으며 길이가 긴 데이터인 경우 RNN보다 순차적 의존성을 효과적으로 학습하여 좋은 성능을 보인다.

LSTM[2]은 순서 정보가 내재되어 있는 데이터들을 처리하는데 적합한 모델로, 나열된 토큰들이 순차적으로 모델에 입력되며 현재 시점에 입력된 토큰은 이전 시점에 입력된 토큰들에 영향을 받아 학습된다.

1.2 GPT-2

GPT-2[3]란 Transformer 디코더를 사용하는 self-attention 기반의 사전학습 및 전이학습 딥러닝 언어 모델이다.

GPT-2의 토큰들은 정해진 크기의 embedding 벡터로 입력되는데 토큰의 위치에 따라 각각 다른 동일한 크기의 positional embedding 벡터를 계산한 후 각각 다른 벡터를 토큰의 위치마다 만들어낸다[2]. 따라서 긴 길이에도 상대적인 값을 넣을 수 있다는 장점을 갖고 있다.

III. The Proposed Scheme

1. 데이터셋 구성

본 실험의 학습 데이터는 요기요 웹 페이지의 치킨 음식점 리뷰 데이터를 수집하여 구축하였다. 총 126,800개의 데이터를 수집한

후 중복 데이터와 치킨 음식점 리뷰를 제외한 나머지 리뷰들을 제거하였다. 100자 이상

긴 문장을 넣을 시 과적합이 발생하여, 글자 수 25자 ~ 100자 사이의 리뷰를 사용하였다.

2. Keras를 이용한 LSTM 문장생성

본 모델은 Seq2Seq를 이용한 LSTM 기반 모형으로 입력 sequence에 알맞은 길이의 sequence를 출력해주는 모델을 적용하였다. Seq2Seq는 인코더와 디코더 두 개의 LSTM으로 구성되며, 인코더 입력 문장만을 사용하고 디코더의 출력을 입력으로 넣는다. 학습 데이터셋은 (입력문장, 출력문장) 쌍으로 구성 되어있고 총 3,509개이다. 또한, 기존 LSTM 모델과 비교하기 위해 Bahdanau et al.(2014)에서 제시한 Attention기법을 적용한 Seq2Seq 모델[4]의 실험도 추가로 진행하였다.

3. KoGPT-2를 이용한 문장생성

KoGPT-2는 위키, 뉴스 등 한국어 원시문장을 사전학습(pre-training)하여 GPT-2의 한국어 성능 한계를 개선한 모델이다. 리뷰 자동생성을 위해 미세조정(fine-tuning)을 위한 추가 학습을 진행하였다. 본 모델에서는 Transformer형 디코더 층을 12층으로 설정하고, BPE 기반의 SentencePiece 토큰화 방식을 사용하여 단어 사전을 생성하였다. 총 학습데이터는 17,067개이다. 스페셜 토큰을 지정하면 맛, 가격, 양, 배달과 관련된 데이터로 미세조정된 KoGPT-2에 의해서 리뷰가 생성된다.

IV. 실험결과

1. Keras LSTM 모델

학습 epoch은 50, batch size는 64로 50번 반복하여 실험한 결과 훈련 정확도는 98.5241%를 기록했지만 검증 정확도는 평균 69.3663%로 목표했던 80%대에는 도달하지 못하였다. 검증 손실 또한 훈련 손실에 비해 큰 결과를 보였다. 특정 반복 이후 부터는 과적합이 발생해 성능이 나아지지 않았는데 조건을 다르게 하여 실험한 결과도 비슷하였다. 이는 데이터의 양이 부족한 탓에 일반적인 데이터 패턴을 학습하여 나타난 결과로 보인다.

Table 1. Keras LSTM 리뷰 생성 예시

Input	Output
가성비 굿	가성비 굿 가격 굿입니다.
배달 너무 느려요	배달이 너무 느려요 조금 실망

2. Bahdanau Attention을 적용한 Seq2Seq 모델

Keras LSTM 모델과의 비교를 위해 학습 epoch은 100, batch size는 2로 학습을 진행한 결과, 훈련 정확도와 손실 모두 Keras LSTM보다 좋지 않았으나 검증 정확도는 90.18%로 더 높은 결과를 보였다. 생성된 결과를 봤을 때 훈련 정확도가 높지 않아 잘못된 리뷰들이 생성된 것으로 보인다.

Table 2. Seq2Seq 리뷰 생성 예시

Input	Output
맛없어요	솔직히 이 너무 맛있어요

3. KoGPT-2 모델

KoGPT-2 모델은 Keras LSTM 모델보다 정확한 리뷰를 생성하였다. 잘못 생성된 리뷰의 경우에는 전혀 다른 의미를 갖는 리뷰를 생성하는 것은 아니지만 뒤에 어색한 문장이 연결되었다. 학습 epoch을 다르게 하였을 때 시험정확도는 93.68%를 넘었지만 , 훈련정확도는 90%를 넘지 못했다. 이는 KoGPT-2 모델 또한 학습 데이터의 양을 증가시키는 것이 필요하다고 생각된다.

Table 3. KoGPT-2 리뷰 생성 예시

Category	Input	Output
맛	매워요	매워요 덜 매우면 엄청 맛있을 거 같아요
양	푸짐	푸짐하고 혼자 먹기에 매우 양이 많습니다

V. Conclusions

현재의 데이터를 이용하여 실험한 결과, 가장 적합한 모델은 KoGPT-2이다. 제안한 모델들 모두 충분한 데이터를 만들지 못하였다는 아쉬움이 있기 때문에 추가적인 데이터 셋을 구축하여 과적합 문제를 해결한다면 지금보다 더 뛰어난 성능이 나올 것으로 예상하고 있다.

REFERENCES

[1] Y. Kim, Y. Hwang, T. Kang, and K. Jung, “LSTM Language Model Based Korean Sentence Generation,” The Journal of Korean Institute of Communications and Information Sciences, vol. 41, no. 5, pp. 592-601, May 2016.

[2] Sunghwan Son, “Category text generation by using GPT-2 model”, Department of Computer Science Graduate School, Kookmin University, Seoul 2020.

[3] Alec Radford, Jefferey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, “Language Models are Unsupervised Multitask Learners,” OpenAI, Feb 2019.

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, “Neural machine translation by jointly learning to align and translate,” CoRR, abs/1409.0473, 2014