LSTM 기반 Sequence-to-Sequence을 활용한 한국어 제목 생성

Korean Title Generation with Sequence-to-Sequence of LSTM

저자
(Authors)

지규빈, 나요셉, 곽경민, 최태영
Kyu Bin Ji, Yoseph Na, Kyung Min Kwak, Tae-Young Choe

출처
(Source)

Proceedings of KIIT Conference , 2020.10, 308-311(4 pages)

발행처
(Publisher)

한국정보기술학회
Korean Institute of Information Technology

URL

http://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE10490812

APA Style

지규빈, 나요셉, 곽경민, 최태영 (2020). LSTM 기반 Sequence-to-Sequence을 활용한 한국어 제목 생성. Proceedings of KIIT Conference, 308-311

LSTM 기반 Sequence-to-Sequence을 활용한 한국어 제목 생성

지규빈*, 나요셉*, 곽경민*, 최태영*

Korean Title Generation with Sequence-to-Sequence of LSTM

Kyu Bin Ji*, Yoseph Na* , Kyung Min Kwak*, and Tae-Young Choe**

요 약

본 연구에서는 LSTM을 기반으로 하는 Sequence-to-Sequence 모델을 활용하여 본문의 단어 간의 유사도를 검출하고 학습하여 추상적 요약을 통해 제목을 생성하는 모델을 제안하였다. 데이터에 단어별 품사를 태깅하고 특수문자를 처리하는 등의 전처리를 수행하였다. 기사의 본 제목과 모델이 생성한 제목의 BLEU 지수를 계산했을 때, 0.558이 가장 높은 결과였음을 확인할 수 있었다.

Abstract

In this paper, we propose a technique that using the Sequence-to-Sequence of LSTM detects and learns similarities between words in the text and generating a title through an abstract summary. Pre-Processing is performed such as tagging the parts by word and processing special characters in the data. When calculating the BLEU score for this title of the article and the title generated by the model, the highest result we have identified is 0.558.

Key words
LSTM, Sequence-to-Sequence, crawling, and deep learning

Ⅰ. 서 론

제목은 작품의 첫인상을 결정하며, 잘못된 제목(선정적이거나, 과장된 제목)으로 인하여 본문의 내용에 대한 이해를 어렵게 한다는 문제점[1]을 야기할 수도 있기 때문에, 제목은 가장 짧은 형태로 본문의 내용을 잘 요약해주어야 한다. 하지만 알맞은 제목을 쓰기란 쉽지가 않은 일이며 이를 위해 많은 시간을 투자하게 된다.

본 연구에서는 알맞은 제목을 작성하기 위하여 LSTM[2] 알고리즘을 기반으로 하는 제목 생성 모델을 제안한다. 특히 Sequence-to-Sequence[3] 모델을 활용하여 본문의 단어 간의 유사도를 검출하고 학습하여, 단순히 문장을 추출하는 것이 아닌 추상적 요약을 하여 제목을 생성하는 모델을 제안한다.

Ⅱ. 관련 연구

알맞은 제목을 작성하기 위한 연구는 다양하게 진행되고 있다. RNN과 강화 학습을 결합한 제목 생성 모델[4]의 경우 TextRank[5] 알고리즘을 적용해 추출된 요약된 본문을 학습에 사용하였고, 적절한 보상함수를 적용해 반복된 단어가 생성되는 문제를 보안하였다. 이는 문장으로 구성된 의미 있는 제목

* 금오공과대학교 컴퓨터공학과
** 금오공과대학교 교신저자(choety@kumoh.ac.kr)

이 아닌 문서에 나타나는 핵심 키워드를 추출하여 나열한다는 한계가 있다. 연구 또한 인공신경망을 이용하여 제목을 생성하는 연구[6]는 존재하나, 한국어에서의 적용은 힘들다는 한계가 있다.
본 모델에서 사용할 sequence-to-sequence 모델은 두 개의 LSTM 시퀀스를 가지고 있어 입력 부분과 출력 부분의 길이를 다르게 설정할 수 있기 때문에 문서 요약 부분에서 좋은 결과를 보여주고 있다.

III. LSTM 기반 Sequence-to-Sequence을 활용한 한국어 제목 추천 모델

3.1 전처리

본 연구에서 제안하는 제목 생성 모델은 전처리를 필요로 한다. 먼저 크롤러를 사용하여 데이터를 크롤링한다. 크롤링된 데이터에서 본문과 원제목을 추출하여 저장하였다. 저장된 데이터를 하나씩 읽어와 특수문자를 제거하거나 글로 변환하였고 KoNLPy[7]를 이용하여 품사를 태깅하는 토큰화를 수행했다. 다음으로 본문 데이터의 단어를 똑같이 맞추고 원제목 데이터는 단어의 개수만 제한한 후 데이터를 학습하면서 시작과 끝을 알리기 위해 <START>토큰과 <END>토큰을 추가하였다. 이러한 데이터를 Sequence-to-Sequence모델에 적용하여 label을 원제목으로 이용하여 모델을 학습하고 저장하였다.

3.2 모듈 설명

(1) 크롤링

데이터를 수집하기 위해 Beautiful Soup[8] 크롤러를 사용하여 데이터를 크롤링한다. 크롤링한 데이터는 네이버 기사의 경제 카테고리에 있는 기사이다. seed URL로부터 크롤링을 시작하며, 기사의 제목과 본문만을 추출하여 csv파일에 저장하도록 한다.

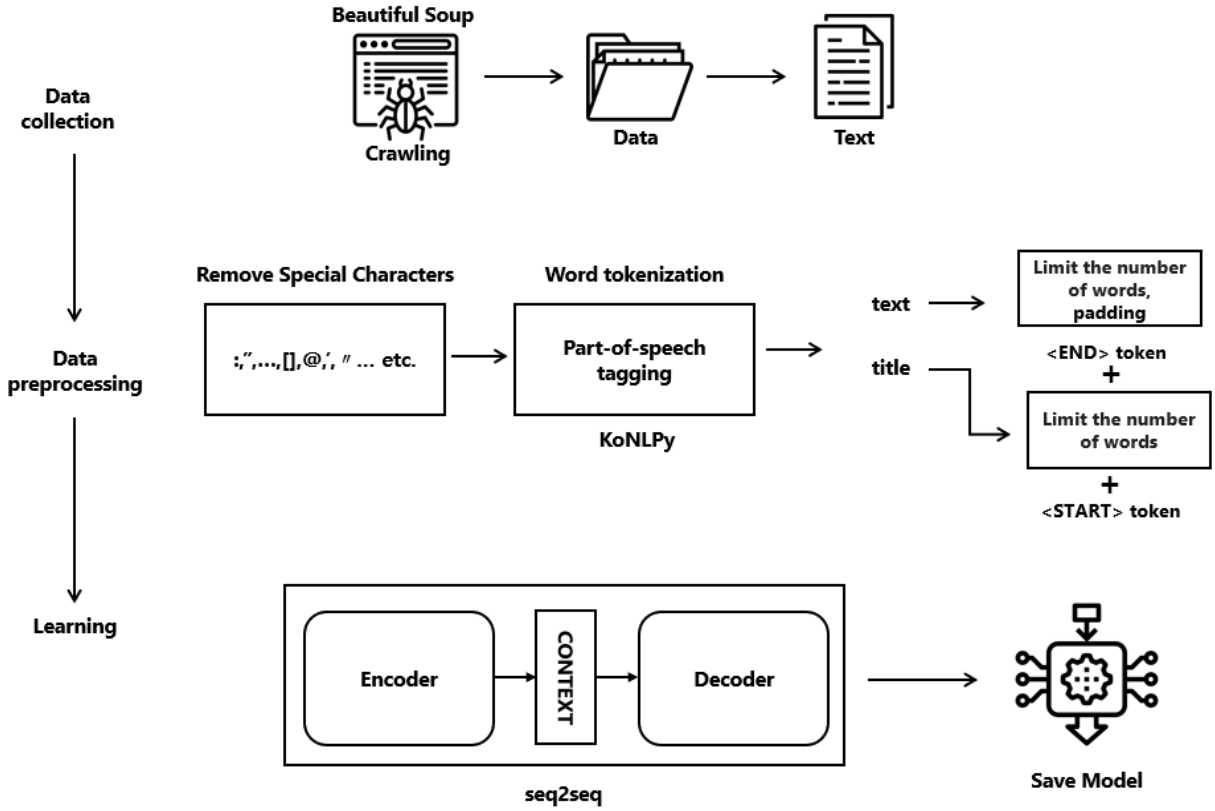

그림 1. 학습 모델 생성
Fig. 1. Make Learning Model

⑵ Sequence-to-Sequence

LSTM셀인 인코더와 디코더로 이루어져 있으며 처음 데이터가 들어오면 본문을 X로 제목을 Y로 나누어 학습한다. 처음 데이터를 임베딩한 후 인코더에 본문이 들어가 다대일 학습을 진행한 후 결과를 디코더로 넣어서 다대다 제목 학습을 수행한다. 이때 마지막에 제목 데이터 수로 Dense layer의 노드수를 결정하고 softmax로 확률을 출력한 후에 loss 함수로 cross entropy error를 이용, nAdam Optimazer를 이용해 학습을 수행한다.

3.3 결과

먼저 2020년 4월 기사를 크롤링하여 약 16000개의 데이터를 모았고, train data를 80%, test data를 20%로 사용되었다.

테스트 데이터 중 약 470개에 BLEU[9] 지수를 계산해보았고 최고점과 최저점의 결과는 표 1과 같다.

Ⅳ. 결 론

본 연구에서 제안하는 제목 생성 모델은 Sequence-to-Sequence 모델을 활용하여 사용자가 입력한 본문의 단어 간의 유사도를 검출하고 학습하여, 단순히 문장을 추출하는 것이 아닌 추상적 요약을 하여 제목을 생성할 수 있었다. 하지만 Sequence-to-Sequence모델을 제목 생성에 유용한 모델이라 채택하였지만, 단어를 연속으로 반복하여 출력할 수도 있다는 모델의 고질적인 문제가 있었다. 보통의 제목은 같은 단어를 연속으로 사용하는 경우는 없기 때문에 생성된 제목에서 중복된 단어를 제거하는 후처리를 해야만 했다.

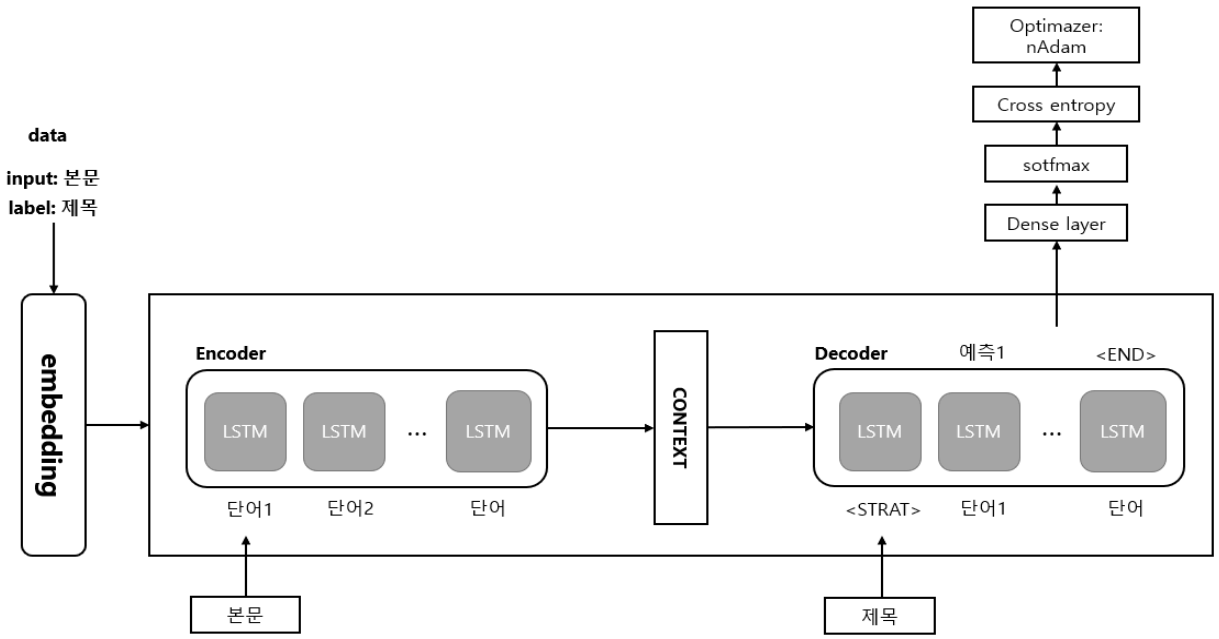

그림 2. Sequence-to-Sequence 모델
Fig. 2. Sequence-to-Sequence model

표 1. BLEU평가 지표 결과
Table 1. Result of BLEU

원제목	생성된 제목	평가지표
신보 국내 최초 사회적경제기업 평가시스템 운영	신 보 국내 최초 사회 적 경 제 기업 평가 시스템 운영 오픈 및 위한 후 실시 확대	0.558
연일 최악의 미국 일자리 지표	美 금융위기 년 만에 만 명 월 말 은행 위 는 또 손해 종합 증권 회 은	0

또한 카테고리를 경제로만 설정하여 수집하였기 때문에 향후 연구에서는 카테고리를 4가지로 늘리고 더 많은 데이터를 수집하여 정확도를 향상시키고자 한다.

참 고 문 헌

[1] 조수선. (2005). 온라인 신문 기사의 제목과 개요 효과. 한국언론학보, 49(2), 5-32.

[2] Merity, Stephen, Nitish Shirish Keskar, and Richard Socher. "Regularizing and optimizing LSTM language models." arXiv preprint arXiv:1708.02182 (2017).

[3] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." Advances in neural information processing systems. 2014.

[4] 조성민, and 김우생. "RNN 과 강화 학습을 이용한 자동 문서 제목 생성." Journal of Information Technology Applications & Management 27.1 (2020): 49-58.

[5] Mihalcea, R. and Tarau, P., “Textrank: Bringing order into text”, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, 2004, pp. 404-411.

[6] Jin, Rong, and Alexander G. Hauptmann. "A new probabilistic model for title generation." COLING 2002: The 19th International Conference on Computational Linguistics. 2002.

[7] Park, Eunjeong L., and Sungzoon Cho. "KoNLPy: Korean natural language processing in Python." Annual Conference on Human and Language Technology. Human and Language Technology, 2014.

[8] Richardson, Leonard. "Beautiful soup documentation." Dosegljivo: https://www. crummy. com/software/BeautifulSoup/bs4/doc/.[Dostopano: 7. 7. 2018] (2007).

[9] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.