텍스트 요약을 위한 스파크 기반 대용량 데이터 전처리

지동준 [1], 전희국 [2], 임동혁 [3]
[1] 광운대학교 인공지능응용학과
[2] ㈜핀다
[3] 광운대학교 정보융합학부
dongjun970@kw.ac.kr,heegook@finda.co.kr, dhim@kw.ac.kr

Spark-Based Big Data Preprocessing for Text Summarization

Dong-Jun Ji[1], Hee-Gook Jun[2], Dong-Hyuk Im[3]
[1]Dept. of Artificial Intelligence Applications, Kwangwoon University
[2]Finda, Seoul, Korea
[3] School of Information Convergence, Kwangwoon University

요 약

텍스트 요약(Text Summarization)은 자연어 처리(NLP) 분야의 주요 작업 중 하나이다. 높은 정확성을 보이는 문서 요약 딥 러닝 모델을 만들기 위해서 대용량 학습 데이터가 필요한데, 대용량 데이터 전처리 과정에서 처리 시간, 메모리 관리 등과 같은 문제가 발생한다. 본 논문에서는 대규모 병렬처리 플랫폼 Apache Spark 를 사용해 추상 요약 딥 러닝 모델의 데이터 전처리 과정을 개선하는 방법을 제안한다. 실험 결과 제안한 방법이 기존 방법보다 데이터 전처리 시간이 개선된 결과를 보이고 있다.

1. 서론

텍스트 요약(Text Summarization)은 자연어 처리(NLP) 분야의 주요 작업 중 하나이다[1]. 텍스트 요약 방식은 크게 추출 요약(Extractive Summarization)과 추상 요약(Abstractive Summarization)으로 분류된다. 추출 요약은 원본 텍스트에서 문장 또는 구의 하위 집합을 가져와 요약을 생성하며, 기존의 문장을 그대로 사용한다[2]. 추상 요약은 원본 문서에 없는 문장 혹은 단어들을 새로 만들어서 새로운 문장을 만들기 때문에 추출 요약보다 모델이 복잡하다[3].

그림 1 은 식료품에 대한 리뷰를 텍스트 요약 모델의 입력 데이터로 사용해서 요약 문장이 생성된 예제를 보이고 있다. 이러한 텍스트 요약 모델이 높은 정확도를 가진 성능을 보장하기 위해서는 대용량 학습 데이터가 필요하며, 그 결과 많은 양의 데이터 전처리 시간, 메모리 관리 등의 문제가 발생한다.

본 논문에서는 텍스트 요약 방법 중 추상 요약 딥 러닝 모델의 데이터 전처리 과정을 개선하는 방법을 제안한다. 대용량 데이터의 신속한 처리를 위해 Apache Spark 기반 병렬 처리 기법[4]을 적용하였으며, 실험을 통해 제안한 방법이 기존 방법 대비 전처리 시간이 개선됨을 보인다.

입력: My cats don't like it. what else can I say to reach 20 words. It was an expensive mistake. My older cat likes the premium edge food for older cats, but she may just be eating it because that is what she gets.
요약: premium edge cat food

입력: These honey sticks are so nice in a cup of tea. You can even remove your tea bag with them!
요약: Lemon Honey Sticks

입력: This taffy is so good. It is very soft and chewy. The flavors are amazing. I would definitely recommend you buying it. Very satisfying!!
요약: Wonderful, tasty taffy

(그림 1) 문서 요약 예제

2. 관련 연구

텍스트 요약 기술은 요약 요인에 따라 분류될 수 있다. 예를 들어, 입력 문서의 수(단일 문서, 다중문서), 문서의 유형(텍스트, 멀티미디어), 출력 유형(추출, 추상)등에 따라 분류할 수 있다[5].

Apache Spark 는 데이터를 병렬로 처리하는 데 적합한 플랫폼이며, 광범위한 데이터를 여러 개의 독립적인 시스템으로 분할한 후에 모든 시스템에서 데이터를 동시에 처리할 수 있다. Apache Spark 는 그래프, 구조화 및 반구조화 데이터 등과 같은 다양한 데이터 구조를 처리할 수 있으며 Scala, Python, Java, R 등 다양한 언어를 지원한다[6].

3. 실험 및 결과

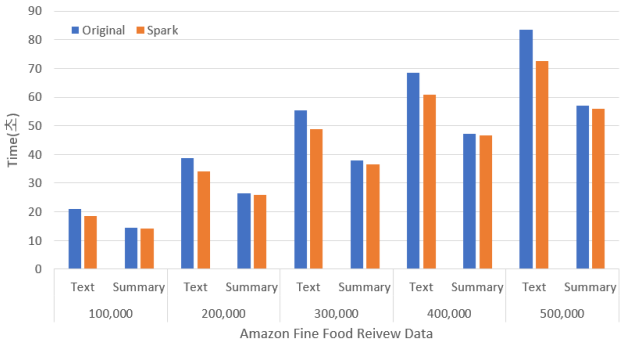
본 연구에서는 Amazon Fine Food Reviews 데이터[7]를 사용하며, jupyter notebook 환경에서 실험을 진행한다. 데이터에는 2012 년 10 월까지 500,000 개의 리뷰 데이터가 존재한다. 그림 1 에서 나온 리뷰와 요약을 통해 실제 요약 작업에 불필요한 단어 및 기호를 제거하여 실험을 진행한다.

기존 데이터 전처리 방법에서 데이터를 50,000 개로 분할하여 전처리 시간을 측정했을 때 입력 리뷰 문장(Text)은 10.7 초, 정답 요약 문장(Summary)은 7.7 초가 소요됐다. 표 1 은 데이터의 수를 50,000 개로 분할시킨 후 파티션을 나누어서 데이터 전처리 시간을 측정한 결과이다. 대용량 데이터를 여러 개의 파티션으로 분할하여 저장할 경우 성능 개선 및 유지보수를 보다 쉽게 할 수 있기 때문에 파티션을 분할하여 다중으로 데이터 전처리를 진행했다.

<표 1> 50,000 개의 데이터를 파티션(Partition)으로 분할한 결과

	Text(초)	Summary(초)
partition=2	9.6206	7.3839
partition=4	9.7075	7.4937
partition=8	9.8051	7.4894
partition=16	9.7177	7.4818
partition=32	9.7048	7.6467

그림 2 는 입력 리뷰 문장(Text)과 리뷰에 대한 정답 요약 문장(Summary)을 기존 방법과 데이터 전처리 속도를 비교한 그래프이다. X 축은 실험 데이터를 100,000 개 단위로 나눈 것을 의미하고, Y 축은 데이터 전처리 속도를 측정한 시간으로 단위는 초(s)를 의미한다. 실험 결과 데이터의 양이 커질수록 제안한 방법의 데이터 전처리 시간이 더 효율적임을 확인할 수 있다.



(그림 2) 기존 방법(Original)과 제안한 방법(Spark) 간의 데이터 전처리 속도 비교

4. 결론

본 연구에서는 자연어 처리 분야의 추상 요약 모델 학습 과정 중 대용량 데이터 전처리 문제를 해결하기 위해 Apache Spark 를 사용한 병렬 처리 기법을 제안하였으며, 실험을 통해 데이터 양이 많아질 수록 이전 방법보다 개선된 처리 속도를 보임을 확인하였다. 향후 연구로 텍스트 추상 요약 분야에서 전처리 뿐만 아니라 모델 학습의 다른 과정에서도 병렬처리 기법을 적용하는 방법을 연구하고자 한다.

Acknowledgement

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.NRF-2021R1F1A1054739).

또한, 본 연구는 과학기술정보통신부 및 정보통신기술진흥센터의 대학 ICT 연구센터지원사업의 연구결과로 수행되었음(IITP-2022-2018-0-01417).

참고문헌

[1] Huang, Y., Feng, X., Feng, X., & Qin, B. (2021). The factual inconsistency problem in abstractive text summarization: A survey. arXiv preprint arXiv:2104.14839.

[2] Moratanch, N., & Chitrakala, S. (2016, March). A survey on abstractive text summarization. In 2016 International Conference on Circuit, power and computing technologies (ICCPCT) (pp. 1-7). IEEE.

[3] Suleiman, D., & Awajan, A. (2020). Deep learning based abstractive text summarization: approaches, datasets, evaluation measures, and challenges. Mathematical problems in engineering, 2020.

[4] Castro, E. P., Chakravarty, S., Williamson, E., Pereira, D. A., & Fox, E. A. (2017, June). Classifying short unstructured data using the Apache Spark platform. In 2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (pp. 1-10). IEEE.

[5] Afantenos, S., Karkaletsis, V., & Stamatopoulos, P. (2005). Summarization from medical documents: a survey. Artificial intelligence in medicine, 33(2), 157-177.

[6] Samiei, S., Joodaki, M., & Ghadiri, N. (2021, May). A scalable pattern mining method using apache spark platform. In 2021 7th International Conference on Web Research (ICWR) (pp. 114-118). IEEE.

[7] McAuley, J. J., & Leskovec, J. (2013, May). From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide Web* (pp. 897-908).