과학기술데이터를 위한 자연어처리 기술 동향

정현지*, 장광선**
*, **한국과학기술정보연구원
hjjeong@kisti.re.kr, gsjang@kisti.re.kr
** 교신저자

Natural Language Processing Trends For Science & Technology Data

Hyun Ji Jeong*, Gwangseon Jang**
*, **Korea Institute of Science and Technology Information
** corresponding author

요 약

연구수행과정에서 발생하는 논문, 특허, 연구보고서 등의 과학기술데이터는 다양한 과학기술지식을 포함한다. 연구자들의 효과적인 연구를 지원하기 위해서는 과학기술데이터 분석을 통한 지식 발견이 필수적이다. 과학기술데이터는 일반 텍스트와는 다르게 다수의 전문용어를 포함하고 있으며, 고유의 양식이 정해져 있고, 텍스트 길이가 대체로 길다는 특징이 있다. 본 고에서는 이러한 과학기술데이터만의 고유한 특징을 반영한 인공지능 기반 자연어처리 기술들을 소개함으로써 과학기술데이터 분석에 대한 이해를 돕고자 한다.

1. 서론

과학기술데이터는 연구 수행 과정에서 발생하는 데이터로 논문, 특허, 연구보고서 등이 있다. 과학기술데이터에는 다양한 과학기술지식이 내재되어 있지만, 지속적으로 생성되는 방대한 과학기술데이터에서 사람들이 필요한 지식을 편리하게 접근하기에 어려움이 있다. 연구자들의 효과적인 연구를 위해 대규모 과학기술데이터를 분석하여 유용한 지식을 발견하는 것은 필수적이다.

텍스트로부터 의미를 이해하고 유용한 정보를 추출하는 자연어처리 기술은 딥러닝 기술의 발전에 따라 최근 몇 년간 비약적인 발전을 이루었다[1,2]. 하지만, 과학기술데이터의 텍스트 정보를 활용한 분석 기술 연구는 아직 시작단계에 불과하다. 본 고에서는 과학기술데이터의 특징을 소개하고, 자연어처리 관련 최근 기술 동향을 분석하여 과학기술데이터 분석에 대한 이해를 돕고자 한다.

본 고의 구성은 다음과 같다. 2장에서 과학기술데이터에 대한 특징을 정의하고 3장에서 과학기술데이터에 적용할 만한 자연어처리 기술들을 분류하고 소개하도록 한다. 4장에서는 각 기술별 최신 동향에 대해 설명하고 5장에서는 본 고의 결론을 내리도록 한다.

2. 과학기술데이터 특징

과학기술데이터는 논문, 특허, 연구보고서와 같이 전문적인 과학기술정보를 포함한 데이터로 주요 특징은 3가지가 있다.

첫 번째, 과학기술데이터에는 다수의 전문용어가 포함되어 있다. 기존 대부분의 자연어처리 모델은 일상용어를 활용하는 응용에 초점을 맞춰 연구해왔다. 따라서, 다수의 전문용어로 구성된 과학기술데이터에 적합한 텍스트 분석 모델이 필요하다.

두 번째 특징은 과학기술데이터는 고유의 양식이 정해져 있는 텍스트 데이터라는 것이다. 논문, 특허, 연구보고서는 출판사에서 지정한 고유의 양식이 있고 저자들은 해당 양식에 맞춰서 문서를 작성한다. 따라서 과학기술데이터는 파싱(Parsing)이 일반 비정형 텍스트에 비해 상대적으로 쉬운 특징이 있다.

마지막으로, 대부분의 과학기술데이터의 텍스트는 길이가 길다. 논문, 특허, 연구보고서는 대부분 수천 단어 이상이며 많게는 수만 단어로 구성된 경우도 있다. 반면, 일반적으로 문서요약에 많이 사용되는 위키피디아 데이터의 경우 평균 수 백개의 단어로 구성되어 있다. 문서의 길이가 길다는 것은 더 복잡한 정보를 포함하고 있다는 것이고 이는 더 복잡한 모델 설계가 요구된다는 것을 의미한다. 따라서 길고 복잡한 과학기술데이터의 특성을 고려한 자연어처리 모델의 연구가 필요하다.

3. 자연어처리 기술 개요

자연어처리란 사람이 사용하는 언어의 의미와 문법적 구조를 파악하여 기계가 이해할 수 있는 표현으로 변환하는 기술이며, 더 나아가 기계가 자연어를 활용하여 텍스트를 생성하는 기술을 포함한다. 본 고에서는 다양한 자연어처리 기술 중에 과학기술데이터를 활용하기 위한 핵심 기술인 개체명 인식(Named Entity Recognition), 사전학습 언어모델(Pretrained Language Model), 문서 요약(Document Summarization), 범용지식질의응답(Open Domain Question Answering)에 대한 최신 연구동향을 소개하도록 한다.

개체명인식이란 비정형 텍스트에서 이름을 나타내는 개체의 위치를 파악하고 해당 개체를 사전 정의한 유형으로 알맞게 분류하는 기술이다. 예를 들어, “영희는 대전광역시에 산다.”라는 문장에서 영희와 대전광역시를 이름을 나타내는 개체로 판단하고, 영희는 ‘사람’으로 대전광역시는 ‘지역’으로 분류하여 태깅하는 기술이다. 예시 문장의 개체명 인식결과는 “영희[사람]는 대전광역시[지역]에 산다.”가 된다. 이러한 개체명 인식은 과학기술데이터에서 실험 데이터 파악, 기술명 파악 등에 활용 가능하다.

사전학습 언어모델은 대용량의 텍스트 데이터를 기반으로 사전에 언어 모델을 훈련하여 언어의 문맥적 의미를 파악함으로써 텍스트 분류, 번역, 요약 등과 같은 다양한 다운스트림 태스크(downstream task)에서 좋은 성능을 내는 모델이다. 사전 학습 모델은 기존에 태스크에 특정된(task-specific) 모델을 생성하는 방식에 비해 비약적인 성능 향상을 이뤄 최근 자연어처리의 대세 기술로 자리매김하였다. 이러한 사전학습 언어모델은 과학기술데이터를 활용한 논문 요약, 분야 분류, 번역 등 다양한 다운스트림 태스크에도 적용가능하다.

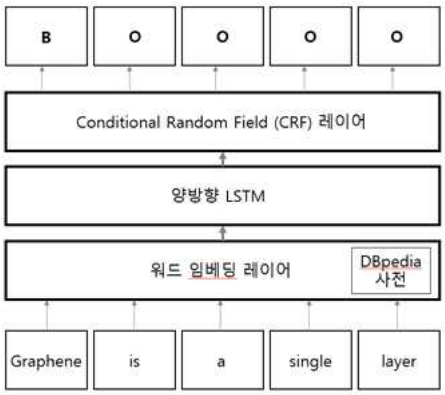
문서 요약은 문서를 입력으로 받아 문서의 내재된 의미를 나타내는 요약된 문장을 출력으로 내는 것을 말한다. 문서 요약 기술은 추출요약과 생성요약 두 가지 방식으로 분류될 수 있다. 추출요약은 문서에 있는 중요 문장을 추출하여 이를 조합하여 요약문을 생성하는 방식이고, 생성요약은 대상 문서에 존재하지 않는 문장들을 활용하여 문서의 내용을 대표하는 요약문을 생성하는 자연어생성(natural language generation) 방식이다. 추출요약은 원문에 있는 문장을 활용하기 때문에 신뢰도가 높으나, 가독성을 확보하기 어렵다. 반면, 생성요약은 요약문은 자연스러우나 의미가 중복된 문장이 반복되는 경우가 많고 구현상 어려움이 있다. 과학기술데이터를 활용한 문서 요약 기술의 대표적인 응용으로는 논문 요약이 있다.

범용질의응답(Open Domain Question Answering, ODQA)은 사용자가 자연어 질의를 입력했을 때, 대규모의 문서 집합으로부터 다양한 주제에 대한 답변을 출력하는 태스크이다. 과학기술데이터에는 다양한 과학기술데이터 관련 지식이 존재하고 연구자들은 데이터가 방대하기 때문에 원하는 정보를 찾기 쉽지 않다. 이러한 환경에서 범용질의응답 기술을 활용한다면 과학기술데이터로부터 사용자가 원하는 답변을 쉽게 찾아줄 수 있고 연구자의 편의성을 증대시킬 수 있다.

4. 과학기술데이터 관련 자연어처리 기술 동향

4.1. SciNER[3]

SciNER는 양방향 LSTM 네트워크와 CRF(Conditional Random Fields)를 기반으로 다수의 워드 임베딩 모델과 함께 외부 지식 소스로서 DBpedia 사전을 사용하여 과학기술데이터 대상 개체명 인식 성능을 향상한 연구이다.



(그림 1) SciNER 구조: 단어 임베딩+양방향 LSTM+CRF. 출처: Reprinted from https://aclanthology.org/D19-1371/, CC BY 4.0

그림 1과 같이 양방향 LSTM은 전향(Forward)의 정보와 후향(Backward)의 정보를 동시에 고려하여 은닉층을 갱신함으로써 단어의 앞, 뒤 관계를 반영하는 모델이다. CRF는 Conditional Random Field의 약자로서 문장 내 형태소의 라벨 간 인접성에 대한 정보를 바탕으로 다음 라벨을 예측하는 모델이다. 양방향 LSTM을 통하여 주어진 문장에 대해 각각 전향, 후향으로 문장의 정보를 고려하여 은닉 상태를 계산하고 CRF를 통해 계산된 은닉 상태에 대해 조건부 확률을 계산하여 개체명을 예측한다.

특성이 다른 두 도메인인 SciNER은 Macromolecules 저널을 기반으로 한 화학 도메인과 Inter-university Consortium for Political and Social Research(ICPSR)를 기반으로 한 사회과학 도메인에서 우수한 성능을 보였다. 첫 번째 Macromolecules 저널 데이터는 Macromolecules

저널의 100개 재료과학 분야 논문을 대상으로 하며, 두 번째 ICPSR 데이터는 ICPSR에서 제공하는 6,368개의 사회과학 분야를 대상으로 한다.

최근에는 BERT 모델에 N개의 레이어를 추가하여 개체명을 태깅하도록 설계한 방법이 양방향 LSTM과 CRF를 결합한 방법보다 높은 성능을 보인다. 그러나 양방향 LSTM과 CRF를 활용한 모델이 추론(Inference)시 건당 처리속도가 BERT 기반 모델 대비 평균 3배 이상 빠른 장점이 있다. 이에 따라 대용량의 과학기술데이터를 실시간성 처리가 필요한 경우는 여전히 양방향 LSTM과 CRF를 결합한 SciNER가 효과적이다.

4.2. SciBERT

SciBERT[4]는 2019년 앨런(Allen) 인공지능 연구소에서 발표한 다분야 과학텍스트(scientific text)를 위한 전문화된 사전학습 언어모델이다. SciBERT는 대규모 과학기술분야 데이터를 수집하여 모델을 훈련하고 각 도메인별 파인튜닝을 수행하여 도메인별 기존 모델 대비 효과를 분석하였다.

SciBERT는 Semantic Scholar의 114만개 논문의 전문을 사용하여 과학텍스트 코퍼스(corpus) 구축하고 해당 코퍼스에서 3만개의 과학분야 단어집합을 새로 생성하였다. 이는 18%는 컴퓨터 과학 도메인, 82%는 바이오의학 도메인 데이터로 구성된다. 대표적인 사전학습 언어모델인 BERT를 그대로 활용하여 훈련하였고 각 도메인 데이터에 적합한 파인튜닝을 하였다.

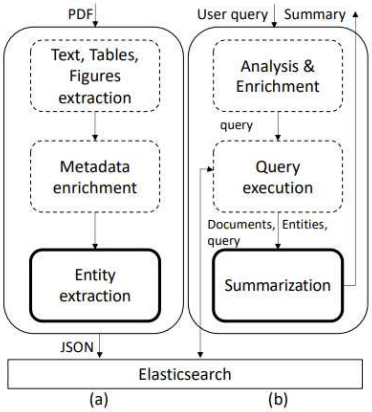
실험은 개체명인식(Named Entity Recognition), 시퀀스 라벨링(PICO extraction), 텍스트 분류(text classification), 관계 분류(relation classification), 의존 구문 분석(dependency parsing) 등 총 5개의 태스크(Task)에 대해 수행하였다. 바이오분야의 BC5CDR[5] 데이터에 대한 개체명 인식 태스크에서 BERT대비 약 3.7%의 성능 향상을 이뤘으며, 컴퓨터분야의 경우 SciERC[6] 데이터에서 약 3.5%의 성능 향상을 이뤘다.

4.3. IBM Science Summarizer

IBM Science Summarizer는 2019년 IBM에서 공개하였으며, 논문 전문의 요약 시스템을 제안한 논문이다[7]. 본 논문의 주요 개선 내용은 길이가 긴 문서에 대한 요약 모델을 설계한 것, 연구자들이 원하는 특정 주제에 대한 요약을 결과로 준다는 것이다.

IBM Science Summarizer의 프레임워크는 그림 2와 같다. 논문을 수집해서 가공하고 인덱싱하는 ingestion pipeline (그림 2의 a)와 사용자의 질의를 입력으로 받아 관련 논문을 검색하고 요약 정보를 제공하는 Summarization부분(그림 2의 b)으로 구성되어 있다. Ingestion pipeline은 먼저 PDF형태의 논문을 파싱(parsing)하여 텍스트, 테이블, 그림을 추출하고 메타데이터를 구성한다. 다음으로, 논문이 어떤 태스크(task)를 대상으로 하는지, 어떤 데이터셋(dataset)을 활용하여 실험을 수행하고, 어떤 메트릭(metric)을 사용하여 성능을 측정했는지를 추출하는 엔티티(entity) 추출 작업을 수행한다. 엔티티 추출 작업은 paperwithcode 사이트의 메타데이터를 활용하여 직접 구축하기도 하고 텍스트 함의(texutal entailment) 기술을 활용하여 자동으로 추출하기도 한다.

Summarization 부분에서는 먼저 사용자 질의를 확장(expansion)하여 문서를 검색하고 문장 분할, 토큰화, 불용어 제거 등의 질의문 전처리 작업을 수행한다. 그리고 질의에 관련된 문서를 검색하여 검색된 문서를 대상으로 요약을 수행한다. 요약은 긴 문서 요약에 장점이 있는 추출 요약(extractive summarization)을 수행한다. 먼저 질의문과 관련된 논문의 섹션을 검색하고, BERT를 활용하여 섹션에 존재하는 문장들의 임베딩 벡터를 생성한다. 다음으로 질의문과 관련 있는 문장 간의 Cross entropy loss는 최소가 되고 질의문과 관련 없는 문장 간의 cross entropy loss는 최대가 되도록 하는 모델을 활용하여 검색된 문장별로 질의문과의 관련도를 계산한다. 그리고 문장 간의 관련도가 높은 문장들을 활용하여 요약문을 생성한다.



(그림 2) IBM Science Summarizer 프레임워크.

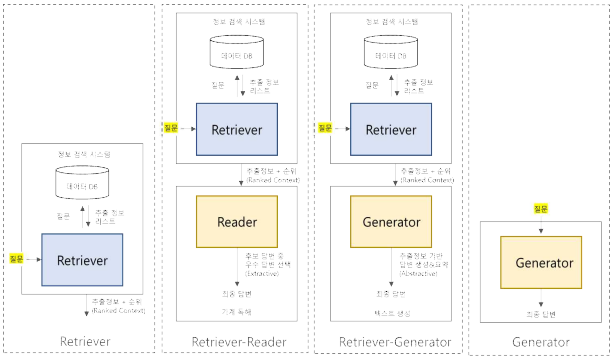
출처: Reprinted from https://aclanthology.org/D19-3036/, CC BY 4.0

IBM Science Summarizer는 12명의 저자에게 24개의 논문을 대상으로 10개의 문장으로 구성된 요약본을 제공하여 설문조사를 통해 성능을 평가하였다. 5점 만점의 설문조사를 하였을 때, 섹션 기반의 논문 요약의 경우 평균 3.32점(표준편차 0.53)의 성능을 보였다.

4.4. GPT-3

GPT-3[1]는 3000억 개로 구성된 데이터셋으로 750억 개의 매개변수를 학습한 대규모 자기 회귀 언어모델로서, openAI가 만든 언어 예측 모델이다. GPT-3는 번역, 글짓

기, 각종 언어 관련 문제 풀이, 사칙연산, 문장으로 웹 코딩 등 다양한 자연어처리 등 다양한 태스크에서 매우 높은 성능을 보였다. GPT-3는 번역, 요약 등 텍스트 생성 기능을 통하여 여러 분야에서 다양한 활용 가능성이 입증되었고, 특히 범용질의응답 분야에서도 우수한 성능을 보였다.



(그림 3) 범용질의응답 분류: Retriever, Retriever-Reader, Retriever-Generator, Generator 구조

범용질의응답은 그림 3과 같이 Retriever, Retriever-Reader, Retriever-Generator, Generator 구조로 크게 4가지로 구분할 수 있다[8]. Retriever 구조는 질문을 검색엔진을 통하여 관련된 정보를 추출하는 검색 위주로 질의응답 시스템보다는 정보 검색 시스템에 가깝다. Retriever-reader 구조는 외부 데이터 소스에서 관련된 정보를 검색하는 Retriever 단계와 검색된 정보에서 해당하는 답변을 정확히 추출하는 Reader 단계로 구성되어 총 2단계로 이루어져 있다. Retriever-Generator는 Retriever 단계에서 검색된 정보를 기반으로 답변을 새롭게 생성하는 방식이다. Generator 구조는 대규모 사전학습 언어모델을 활용한 방법으로 대규모의 파라미터를 활용하여 사전학습한 대규모의 비지도 텍스트 정보를 기억하는 방식이다. 따라서 전에 언급한 3가지 구조와는 달리 질문에 대한 답변을 생성할 때, 외부 데이터를 활용하지 않는다. GPT-3가 Generator 구조에 해당하는 대표적인 모델이라고 할 수 있다. GPT-3 등장 이전까지는 대부분의 질의응답 분야는 외부 데이터 소스를 활용하는 Retriever-Reader 구조와 Retriever-Generator 구조가 주로 연구 및 활용되었으나, GPT-3 등장 이후 대규모 언어모델을 통한 질의응답 시스템을 구축하고자 하는 시도가 늘어나고 있다.

GPT-3는 TrivaQA 데이터를 대상으로 한 질의응답 태스크에서 퓨샷러닝(Few-shot Learning)만으로 기존 최고 성능 모델의 성능을 뛰어넘었다. 과학기술데이터 분야에서 주로 현재 Retriever-Reader 구조기반으로 질의응답 시스템이 구축하여 활용 중이며, Retriever-Generator 구조 기반의 연구도 활발하게 진행 중이다. 그러나 GPT-3와 T5[2] 같은 대규모 모델을 기반으로 하여 외부 데이터 참조 없이 ClosedBookQA 방식의 질의응답 방식이 우수한 성능을 보임에 따라, 대규모 모델을 과학기술데이터 분야 질의응답에 적용한다면 더 높은 성능을 보일 수 있을 것이라 기대된다.

5. 결론

본 고에서는 과학기술데이터 관련 자연어처리 기술 동향에 대해 살펴보았다. 과학기술데이터는 일반 데이터와는 다르게 다수의 전문용어가 포함되어 있고 길이가 긴 문서이며 정형화된 양식이 존재한다는 특징이 존재한다. 이러한 과학기술데이터 특징을 반영한 자연어처리 기술 연구가 필수적이며, 본 고에서는 과학기술데이터 분석 시 많이 활용되는 개체명 인식, 사전학습 언어모델, 문서 요약, 범용질의응답 기술의 동향을 살펴보았다.

과학기술데이터에 내재 된 다양한 과학기술 지식을 추출하기 위해 몇몇 연구가 진행되고 있지만, 아직 초기 단계에 머물고 있다. 앞으로 과학기술데이터 고유의 특징을 잘 반영하는 과학기술데이터 전용 자연어처리 모델 연구가 확산될 것이며, 이는 연구자들의 편의성 도모에 기여할 것으로 생각된다.

사사

본 연구는 (2021년도) 한국과학기술정보연구원(KISTI) 기본사업 과제로 수행한 것입니다.

참고문헌

[1] Brown, Tom B., et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020)

[2] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." arXiv preprint arXiv:1910.10683

[3] Hong, Zhi, et al. "SciNER: extracting named entities from scientific literature." International Conference on Computational Science. Springer, Cham, (2020).

[4] Iz Beltagy, et al., "SciBERT: A Pretrained Language Model for Scientific Text", EMNLP-IJCNLP, (2019).

[5] Jiao Li, et al., "BioCreative V CDR task corpus: a resource for chemical disease relation extraction. Database", The journal of biological databases and curation (2016).

[6] Yi Luan, et al.,. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In EMNLP, (2018)

[7] Shai Erera., et al. "A Summarization System for Scientific Documents", EMNLP-IJCNLP, (2019).

[8] Lewis, Patrick, et al.. "Question and answer test-train overlap in open-domain question answering datasets." arXiv preprint arXiv:2008.02637, (2020).