학술대회 및 저널별 기술 핵심구 추출 모델

정현지* · 장광선 · 김태현 · 신동구

한국과학기술정보연구원

A Keyphrase Extraction Model for Each Conference or Journal

Hyun Ji Jeong · Gwangseon Jang · Tae Hyun Kim · Donggu Sin

Korea Institute of Science and Technology Information

E-mail : hjjeong@kisti.re.kr / gsjang@kisti.re.kr / heemang@kisti.re.kr / lovesin@kisti.re.kr

요 약

연구 동향을 파악하는 것은 연구 수행 시 필수적인 요소이다. 대부분의 연구자들은 관심분야의 학술대회 및 저널을 대표하는 기술 핵심구나 관심 분야를 검색함으로써 연구 동향을 파악한다. 하지만, 최근 인공지능과 같은 특정 분야의 경우 한 개의 학술대회에 한 해당 수백~수천 개의 논문이 출간되기 때문에 전체 분야의 경향성을 파악하는 데 어려움이 존재한다. 본 논문에서는 학술대회 또는 저널 제목을 활용하여 기술 핵심구를 자동으로 추출함으로써 연도별 학술대회 및 저널의 연구 동향 파악을 지원하고자 한다. 핵심구 추출은 문장 또는 문서를 대표하는 주요 구문을 추출하는 작업으로서 검색, 요약, 내용 파악 등을 위해 근간이 되는 기술이다. 기존 사전학습 언어모델 기반의 핵심구 추출 모델은 문서 단위의 긴 텍스트를 기준으로 모델링 하였기 때문에 제목 단위의 짧은 텍스트에서는 성능이 낮아진다는 단점이 존재한다. 본 논문에서는 짧은 텍스트에 강인하면서 단어 자체의 중요도를 고려한 학술대회 및 저널의 기술 핵심구 추출 모델을 제안하고자 한다.

ABSTRACT

Understanding research trends is necessary to select research topics and explore related works. Most researchers search representative keywords of interesting domains or technologies to understand research trends. However some conferences in artificial intelligence or data mining fields recently publish hundreds to thousands of papers for each year. It makes difficult for researchers to understand research trend of interesting domains. In our paper, we propose an automatic technology keyphrase extraction method to support researcher to understand research trend for each conference or journal. Keyphrase extraction that extracts important terms or phrases from a text, is a fundamental technology for a natural language processing such as summarization or searching, etc. Previous keyphrase extraction technologies based on pretrained language model extract keyphrases from long texts so performances are degraded in short texts like titles of papers. In this paper, we propose a techonolgy keyphrase extraction model that is robust in short text and considers the importance of the word.

키워드

Keyphrase extraction, Natural language processing, Academic data analysis, Pretrained language model

Ⅰ. 서 론

하루에도 수천 개 이상의 논문이 쏟아지는 논문의 홍수 시대에 연구자들이 관심 분야의 연구 동향을 제때에 파악하는 것은 쉽지 않다. 특히, 인공지능 분야와 같은 특정 분야는 우수학술대회 1개에서 1년에 출간되는 논문만 수천 건에 이르기 때문에 연구자들이 매년 많이 연구되는 기술의 명칭조차 파악하는 것도 어려운 일이 되었다. 본 논문에서는 연구자들의 연구 동향 파악을 지원하기 위

* speaker

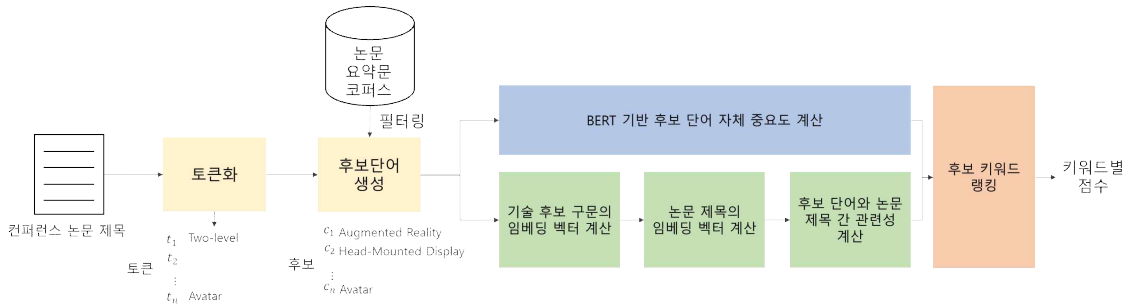

그림 1. 기술 핵심구 추출 모델 개념도

해 저널 또는 학술대회에서 출간되는 논문 제목들을 활용하여 각 저널 및 학술대회를 대표하는 핵심구(keyphrase)를 추출하는 기술을 제안한다.

핵심구 추출 기술은 문장 또는 문서를 대표하는 단어 또는 구문을 추출하는 기술로서 지도학습과 비지도 학습 기반 기술로 분류할 수 있다. 본 논문에서는 지속적으로 발간되는 논문들의 핵심구를 레이블 데이터 없이 발견하기 위해서 비지도학습 기반 핵심구 추출 모델을 제안한다.

기존 비지도학습 기반 핵심구 추출 모델은 Yake[1]와 같은 빈도수 기반 모델, TextRank[2]와 같은 그래프 기반 모델 그리고 최근에 많이 연구되고 있는 사전학습 언어모델을 활용한 핵심구 추출 모델[3]로 분류할 수 있다. 빈도수 및 그래프 기반 핵심구 추출 모델은 정답 데이터 없이도 빠르게 핵심구를 추출한다는 장점이 있지만, 단어 자체의 의미를 고려하지 않아 정확도가 감소하는 경향이 있다. 반면, 사전학습 언어모델 기반 핵심구 추출 모델은 대부분 일정 길이 이상의 문서에서 핵심구를 추출하는 것을 목적으로 하기 때문에 짧은 문장에서 핵심구 추출에는 성능 저하가 발생한다. 본 논문에서는 이와 같은 문제점을 해결하기 위해 짧은 문장에서 강인한 비지도학습 기반 핵심구 추출 모델을 제안한다.

II. 기술 핵심구 추출 모델

본 장에서는 저널 및 학술대회별 핵심구 추출을 위한 기술 핵심구 추출 모델에 대해 설명한다. 본 논문은 저널 및 학술대회에서 출간한 논문 제목 집합을 활용하여 여러 제목들에 빈번하게 존재하는 구문을 추출하는 방식으로 저널 및 학술대회의 핵심구를 추출한다. 논문 제목은 해당 논문을 대표하는 주요 키워드를 포함하기 때문에 많은 논문 제목에 존재하는 구문은 해당 저널 및 학술대회를 대표하는 핵심구이다.

그림 1은 본 논문에서 제안하는 기술 핵심구 추출 모델의 개념도를 나타낸다. 저널 및 학술대회별 핵심구 추출을 위해 먼저 연도별로 저널 및 학술대회에서 출간된 논문 제목 데이터를 하나의 문서로 생성한다. 예를 들어, 2019년에 KDD에서 출간된 논문 제목 집합을 하나의 문서로 생성한다. 다음으로 입력 문서를 토큰화하고 특정 문법 규칙에 적합한 구문을 추출하여 후보 단어를 선정한다. 이때, 후보 단어를 단순히 문법 규칙을 기준으로 선정하였기 때문에 특정 논문에서만 사용되는 고유명사처럼 기술 핵심구로서 적합하지 않은 경우들이 있다. 따라서, 사전에 구축한 논문 요약 코퍼스를 활용하여 논문 요약 코퍼스에 일정 빈도 이상 존재하는 구문을 최종 후보 단어로 선정한다.

단어 자체의 중요도를 반영한 핵심구 추출 모델 개발을 위해 그림1의 상단과 같이 사전학습 언어모델인 BERT 훈련 시 계산된 어텐션 값을 활용하여 후보 단어 자체의 중요도를 계산한다. 또한, 짧은 문장인 제목에서도 효과적으로 핵심구를 추출하기 위해 후보 구문별 문장에서의 중요도를 계산한다. 이는 기술 후보 구문의 임베딩 벡터와 논문 제목의 임베딩 벡터 간 코사인 유사도를 통해 계산된다. 기술 후보 구문 및 논문 제목의 임베딩 벡터는 기술 후보 구문 및 논문 제목에 포함된 토큰들의 BERT 임베딩 벡터를 셀프어텐션한 각각의 임베딩 벡터를 활용한다. 마지막으로, 단어 자체의 중요도와 후보 단어의 제목에서의 중요도를 가중합하여 최종적으로 후보 단어별로 우선순위를 매긴다. 우선순위 결과에 따라 상위에 존재하는 후보 단어가 핵심구로 추출된다.

III. 실험

본 장에서는 기술 핵심구 추출 모델의 실험 방법 및 실험 결과를 기술한다. 기술 핵심구 추출 모델의 정확도 평가를 위해 2021년도 IT분야 15개 학술대회의 주요 기술 키워드 학습데이터를 전문가를 활용하여 구축하였다. 학습데이터는 연도별 학술대회에 해당하는 기술 핵심구로 구성된다.

표 1은 기술 핵심구 추출 모델의 비교 실험 결

과를 나타낸다. 제안모델이 기존에 많이 활용되는 핵심구 추출 모델인 TextRank[2]와 Yake[1]보다 대부분의 데이터에서 좋은 성능을 나타냄을 알 수 있다.

표 1. 기술 핵심구 추출 모델 실험 결과

데이터	TextRank	Yake	제안모델
ICDE	0.0	0.0	**0.1**
PODS	0.2	**0.5**	0.2
SIGMOD	0.0	**0.2**	0.1
VLDB	0.0	0.2	**0.3**
CIKM	0.0	0.1	**0.1**
ICDM	0.0	0.1	**0.3**
KDD	0.0	0.2	**0.2**
SIGIR	0.0	**0.3**	0.2
WWW	0.0	0.1	**0.5**
ASPLOS	0.0	0.3	**0.3**
DAC	0.0	0.0	**0.5**
HPCA	0.0	0.0	**0.2**
ISCA	0.0	0.3	**0.4**
ISMAR	0.2	0.2	**0.2**
UIST	0.0	0.2	**0.3**

Ⅴ. 결 론

본 논문에서는 연구자의 연구동향 파악을 지원하기 위해 학술대회 및 저널을 대표하는 핵심구 추출 기법을 제안하였다. 제안한 기법은 사전 학습 언어모델을 활용하여 단어 자체의 의미를 고려하고 셀프어텐션 기술을 활용하여 문장 전체와 관련성이 높은 핵심구를 추출함으로써 짧은 문장에서도 효과적으로 핵심구를 추출하였다. 그 결과, 전문가를 활용하여 구축한 학술대회 데이터에서 제안한 방법이 기존 연구보다 성능이 향상됨을 보일 수 있었다.

Acknowledgement

본 연구는 2022년도 한국과학기술정보연구원(KISTI) 주요사업 과제로 수행한 것입니다(NTIS 과제고유번호 1711173845).

References

[1] Campos, Ricardo, et al. “YAKE! Keyword extraction from single documents using multiple local features.” Information Sciences 509 (2020): 257-289.

[2] Mihalcea, Rada, and Paul Tarau. “Textrank: Bringing order into text.” Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.

[3] Ding, Haoran, and Xiao Luo. “AttentionRank: Unsupervised Keyphrase Extraction using Self and Cross Attentions.” Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.