J. Internet Comput. Serv.
ISSN 1598-0170 (Print) / ISSN 2287-1136 (Online)
http://www.jics.or.kr

생의학 분야 키워드 추출 모델에 대한 비교 연구☆

Comparative Study of Keyword Extraction Models in Biomedical Domain

이 동 희1 권 순 찬1 장 백 철1*
Donghee Lee Soonchan Kwon Beakcheol Jang

요 약

생명 공학 및 의학 분야의 논문 수 증가에 따라 문헌 속에서 중요한 정보를 빠르게 찾아 대응하기 위한 키워드 추출의 필요성이 대두되고 있다. 본 논문에서는 생의학 분야에서의 키워드 추출에 대한 다양한 비지도 학습 기반 모델 및 BERT 기반 모델의 성능을 종합적으로 비교하였다. 실험 결과 생의학 분야에 특화된 데이터로 학습된 BioBERT 모델이 가장 높은 성능을 보였다. 이를 통해 생의학 분야의 키워드 추출 연구에서 적절한 실험 환경을 구성하고 다양한 모델을 비교 분석하여, 향후 연구에 필요한 정확하고 신뢰할 수 있는 정보를 제공하였다. 이뿐만 아니라, 다른 분야에서도 키워드 추출에 대한 비교적인 기준과 유용한 지침을 제공할 수 있을 것이라 기대한다.

☞ 주제어 : 키워드 추출, 자연어처리, 딥러닝, 생의학

ABSTRACT

Given the growing volume of biomedical papers, the ability to efficiently extract keywords has become crucial for accessing and responding to important information in the literature. In this study, we conduct a comprehensive evaluation of different unsupervised learning-based models and BERT-based models for keyword extraction in the biomedical field. Our experimental findings reveal that the BioBERT model, trained on biomedical-specific data, achieves the highest performance. This study offers precise and dependable insights to guide forthcoming research in biomedical keyword extraction. By establishing a well-suited experimental framework and conducting thorough comparisons and analyses of diverse models, we have furnished essential information. Furthermore, we anticipate extending our contributions to other domains by providing comparative experiments and practical guidelines for effective keyword extraction.

☞ keyword : Keyword Extraction, NLP, DeepLearning, Biomedicine

1. 서 론

생명 공학 및 의학 기술의 발전과 COVID-19와 같은 팬데믹으로 인해 관련 분야의 문헌 수가 증가하고 있으며 이에 따라 넘쳐나는 문헌 속에서 중요한 정보를 빠르게 찾아 대응하는 것이 중요해졌다. 키워드 추출(Keyword Extraction)은 문서를 가장 잘 설명하는 단어인 키워드를 식별하고 추출하여 문서의 정보 중 가장 관련성이 높은 정보를 나타내는 작업이며[1] 추출된 키워드를 사용하여 현재 상황에 필요한 정보를 신속하게 획득한 뒤 문제에 대응할 수 있다. 그러나 생명 공학 및 의학 분야의 전문 용어가 자주 사용되는 논문의 특성상 키워드 추출이 상대적으로 어려워 키워드 추출에 관한 연구가 활발히 이루어지고 있지 않다. 기존의 생명 공학 및 의학 분야에 대한 키워드 추출 모델 성능 비교 연구의 경우 비지도 학습 모델을 대상으로 성능을 비교하거나[2], 임베딩 기법을 통한 언어 모델의 성능을 비교하는 연구[3]가 있었지만 종합적인 비교 연구는 없었다.

따라서 본 연구는 생명 공학 및 의학 분야의 키워드 추출에 관한 비지도 학습 및 BERT 기반 모델의 성능을 종합적으로 비교하여 생명 공학 및 의학 분야에서 효과적으로 키워드를 추출할 수 있는 모델에 대한 정보를 제공하고자 한다. 다양한 키워드 추출 모델의 성능을 비교하였으며 실험을 통해 미세 조정된 BERT 기반의 키워드 추출 모델이 미세 조정되지 않은 모델이나 비지도 학습 기반의 키워드 추출 모델보다 높은 성능을 보였다는 것을 확인하였다. 그 중에서도 BioBERT를 미세 조정한 모델의 F1-Score가 0.315로 가장 좋은 성능을 보여주었다.

1 Graduate School of Information, Yonsei University., Seoul, 03722 Korea.
* Corresponding author (bjang@yonsei.ac.kr)
[Received 10 July 2023, Reviewed 20 July 2023, Accepted 11 August 2023]
☆ This work was supported by the Yonsei University Research Fund under Grant 2023-22-0104.

본 연구는 생명 공학 및 의학 분야에서 키워드 추출에 적합한 모델 비교와 함께, 연구의 전반적인 방향을 제시한다. 더불어, 생명 공학 및 의학 분야 외의 다른 분야에서도 키워드 추출에 대한 기준과 지침을 제공하는 역할을 할 것으로 기대된다. 이를 통해 정보 검색, 문서 요약 등의 향후 연구에서 더욱 우수한 결과를 도출해 낼 수 있을 것으로 기대된다.

본 논문은 총 5개의 장으로 구성되어 있으며, 1장에서는 연구의 배경, 필요성, 목적 그리고 기존 연구와의 차이점을 살펴보고 본 연구의 결과를 제시한다. 2장에서는 비지도 학습 기반 키워드 추출 그리고 BERT 기반 키워드 추출과 관련한 연구를 소개한다. 3장에서는 실험에 사용된 데이터 세트, 실험 환경, 성능 평가 방법 그리고 사용한 키워드 추출 모델과 연구 방법을 설명한다. 4장에서는 각 키워드 추출 모델의 성능 측정 결과를 평가하고 실제 추출된 키워드들을 비교한다. 마지막으로 5장에서는 결론과 시사점 및 향후 연구 계획에 관해 설명한다.

2. 관련 연구

2.1 비지도 학습 기반 키워드 추출

비지도 학습 키워드 추출 기법은 크게 통계 기반, 그래프 기반, 임베딩 기반으로 나눌 수 있다. 통계 기반으로는 FirstPhrases, TfIdf등이 있으며 최근에는 키워드 후보들의 로컬 특징을 추출하여 중요도를 계산하는 YAKE[4]등의 모델이 등장하였다.

그래프 기반의 키워드 추출 방법의 시초로는 TextRank[5]를 들 수 있다. TextRank는 구글의 PageRank[6] 알고리즘을 텍스트 문서에 적용시켜 그래프 기반 중요도를 계산하여 중요 단어나 문장을 추출하는 모델이다. SingleRank[7] 는 TextRank에서 처럼 단어와 문장을 그래프의 노드로 표현하지만 주변 단어의 중요도와 연결성을 고려하여 가중치를 반복하여 업데이트 시킨 후 최종적으로 중요도가 가장 높은 문장이나 단어를 키워드로 추출한다는 차이점이 있다. TopicRank[8]에서는 군집화 개념을 추가하였다. 토픽으로 군집화 후 각 토픽별 중요도를 측정해 토픽 내에서 가까운 단어나 문장을 추출하여 키워드로 선정했다. 그리고 MultipartiteRank[10]는 텍스트 문서들의 구성 요소간 관계를 그래프로 표현하여 중요 토픽을 식별한다.

마지막으로 임베딩 기반의 키워드 추출 방법으로는 키워드 추출에 임베딩 모델을 처음 적용한 EmbedRank[9]가 대표적이다. EmbedRank는 단어와 문장을 임베딩하여 벡터화 한 후 유사성을 판단한다. 최근, 비지도학습 기법이지만 임베딩 모델과 사전학습된 언어 모델인 ELMO를 결합한 SifRank[11] 또한 등장하였으며, 현재 2023년 기준 사전학습 언어 모델을 통해 프롬프트를 생성할 확률을 계산하는 방법을 활용한 모델인 PromptRank[12]가 키워드 추출에서 가장 높은 성능을 보여주고 있다.

2.2 BERT 기반 키워드 추출

구글의 트랜스포머를 기반으로 한 대규모 언어 모델인 BERT[13]의 등장으로 다른 자연어처리 작업들이 그렇듯 키워드 추출 작업에서도 BERT 기반 모델들이 효과적인 성능을 보여주고 있다. BERT는 트랜스포머의 인코더 구조를 가져온 양방향 언어 모델로 대량의 데이터로 사전학습을 하고 특정 작업에 맞게 미세 조정 된다. 이를 통해 BERT는 단어들 간 상호작용을 고려하고 문맥 정보를 파악할 수 있어 원문을 더 잘 이해하고 효과적으로 키워드를 생성해낸다.

BERT 이후 꾸준하게 관련 연구가 진행됨에 따라 BERT 모델을 세밀하게 변형하여 성능을 향상시키는 연구나 특정 도메인의 데이터를 사전학습한 형태로 모델을 발전시키는 연구가 등장하였다. RoBERTa[14], AlBERT[15] 등이 대표적인 예로, 이들은 BERT의 아이디어를 기반으로 성능을 향상시켜 다양한 자연어처리 과제들을 해결한다. 도메인 특화 모델로는 Biobert[16], SciBERT[17]등 이 있어 특정 분야에 대한 자연어처리 작업에 효과적으로 사용되고 있다.

3. 연구 방법

3.1 데이터 및 전처리

본 연구는 생명 공학 및 의학 분야 논문의 초록과 키워드 3천 개로 구성된 벤치마크 데이터 세트[2]를 사용했으며, 논문 제목은 키워드 추출 작업에 사용하지 않아 전처리 과정에서 제거하였다. 3천 개의 데이터 중 80%를 사전학습된 모델을 미세 조정하기 위한 학습 데이터로 사용하였고 나머지 20%를 모델의 성능을 비교하기 위한 평가 데이터로 사용하였다. 사전 학습된 모델을 미세 조정하기 전 표 1과 같이 BIO 형식의 라벨[18]로 변환하는 전처리 과정을 진행했다. BIO 형식의 라벨은 개체명 인식이나 토큰 분류 작업에서 일반적으로 사용되며 B는 키워드의 시작, I는 키워드의 시작과 연결된 단어, O는 키워드 이외의 단어를 의미한다.

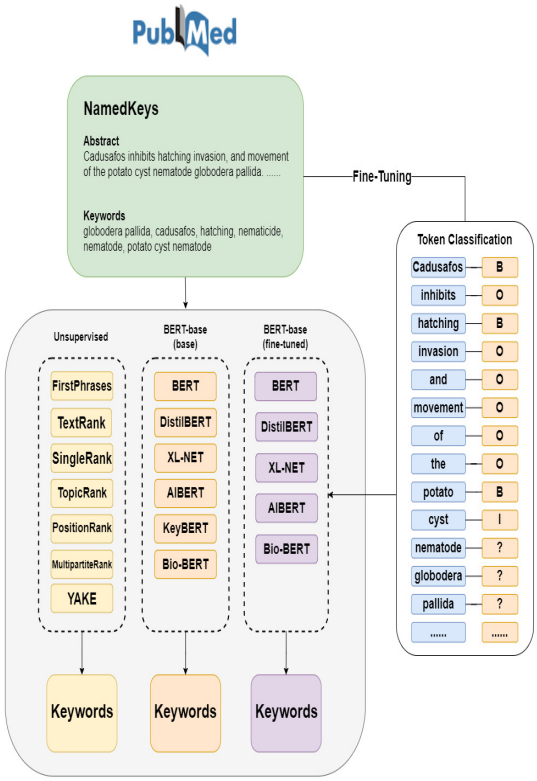
(표 1) 데이터 전처리 예시
(Table 1) Examples of data preprocessing

예시 1	
논문 초록	cadusafos inhibits hatching invasion, and movement of the potato cyst nematode globodera pallida. ..... the inhibition of hatching was permanent.
키워드	globodera pallida, cadusafos, hatching, nematicide, nematode, potato cyst nematode
BIO Label	[B, O, B, O, O, ... O, O, B, O, O]
예시 2	
논문 초록	vera peters and the conservative management of early-stage breast cancer. ..... in the years that followed, prospective randomized studies confirmed her findings.
키워드	vera peters, cancer of the breast, early stage, lumpectomy, radical mastectomy,'survival
BIO Label	[B, I, O, O, O, ... O, O, O, O, O]

3.2 키워드 추출 모델

본 연구는 주어진 생의학 문헌에 대한 키워드 추출 모델의 성능을 종합적으로 비교하기 위해 다양한 비지도 학습 모델과 BERT 기반의 모델을 사용하였다. 이후 비지도 학습 기반 모델, 미세 조정하지 않은 모델, 미세 조정한 모델에 모두 평가 데이터를 사용하여 키워드 추출 성능을 평가하였으며 실험의 전반적인 과정은 그림 1과 같다.

실험에 사용한 비지도 학습 키워드 추출 모델은 통계, 그래프 기반의 모델로 분류할 수 있다. 통계 기반의 비지도 키워드 추출 모델로 FirstPhrases와 YAKE[4]를 사용하였다. 그래프 기반의 비지도 키워드 추출 모델로는 TextRank[5], SingleRank[7], TopicRank[8], PositionRank[19] 그리고 MultipartiteRank[10] 모델을 실험에 사용하였다. BERT 기반의 키워드 추출 모델로는 BERT[13]와 BERT를 변형한 모델인 DistillBERT [20], XL-NET[21], AlBERT[15]를 실험에 활용하였고. 또한 생의학 분야에 특화된 데이터로 사전 학습된 BioBERT[16]와 키워드 추출 작업에 특화된 KeyBERT* 모델도 실험에 사용하였다.



(그림 1) 키워드 추출 모델의 성능 비교 실험 프레임워크
(Figure 1)Performance comparison experiment framework of keyword extraction model

(표 2) BERT 기반 모델의 파라미터 개수 및 미세 조정 학습 시간 비교
(Table 2) Comparing the number of parameters in a BERT-based model and the training time for fine-tuning

Bert 기반 모델	파라미터 개수	학습 시간
BERT	110M	1240sec
DistilBERT	66M	1160sec
XL-NET	117M	600sec
AlBERT	11M	1190sec
KeyBERT	66M	-
Bio-BERT	110M	2750sec

* https://github.com/MaartenGr/KeyBERT

(표 3) 실험에 사용된 하드웨어 및 소프트웨어 환경
(Table 3) Hardware and software environments used in the experiment

구분	내용
CPU	Intel i7-11700K @ 3.60GHz
GPU	Nvidia Geforce RTX 3060
RAM	32GB
OS	Windows 10
Language	Python 3.9.0
DL Framework	Pytorch 1.12.1
CUDA	CUDA 11.3

BERT기반 자연어 생성 모델들의 파라미터 개수는 표 2와 같고 사전 학습된 언어 모델을 특정 작업에 사용하기 위해 모델의 파라미터를 조정하는 미세 조정이 필요하며 키워드 추출은 토큰의 라벨을 분류하는 토큰 분류 작업과 같다. 따라서 본 연구는 BERT 기반의 사전 학습된 언어 모델을 키워드 추출에 적용하기 위해 토큰에 BIO 라벨이 부여된 데이터를 사용하여 모델이 토큰의 라벨을 분류할 수 있도록 미세 조정하였다. BERT, DistillBERT, XL-NET, ALBERT, BioBERT 모델을 미세 조정하였으며 키워드 추출 작업에 특화된 KeyBERT 모델은 미세 조정하지 않았다. 미세 조정에 사용한 하이퍼 파라미터는 epoch = 8, learning rate = 8e-6, batch size = 8, seed = 0이며 Adam optimizer를 사용하였다. BERT 기반 모델들의 미세 조정에 소요된 시간은 표 2와 같고, 실험에 사용된 하드웨어 및 소프트웨어의 자세한 환경은 표 3과 같다.

3.3 평가 지표

키워드 추출 성능을 평가하는데 일반적으로 사용되는 평가 지표인 정밀도(Precision), 재현율(Recall), F1-Score를 사용하여 키워드 추출 모델의 성능을 평가한다. 정밀도, 재현율, F1-Score는 표 4와 같은 혼동행렬을 기준으로 측정하며 정밀도, 재현율, F1-Score를 산정하는 방법은 각 식 (1), (2), (3)과 같다.

(표 4) 혼동행렬 표
(Table 4) Confusion Matrix

	Predict	
Actual	TP	FN
	FP	TN

$$precsion = \frac{TP}{TP + FP} \quad (1)$$

$$recall = \frac{TP}{TP + FN} \quad (2)$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \quad (3)$$

4. 실험 결과

4.1 모델 성능 비교

실험을 통해 다양한 키워드 추출 모델의 성능을 비교한 결과는 표 5와 같으며 P는 정밀도, R은 재현율, F는 F1-Score를 의미한다. 비지도 학습 기반 키워드 추출 모델의 경우 MultipartiteRank의 F1-Score가 0.146으로 가장 높은 성능을 보여주었다. 이외에도 YAKE와 TopicRank의 F1-Score는 각각 0.141, 0.129로 다른 비지도 학습 기반 키워드 추출 모델보다 상대적으로 높은 성능을 보여주었다.

반대로 BERT 기반의 키워드 추출 모델의 경우 미세 조정 과정을 거치지 않았을 때 모두 낮은 성능을 보여주었다. 이는 많은 하위 작업에 적용할 수 있도록 설계된 BERT 모델의 특성상 특정 작업을 수행하도록 미세 조정하지 않았을 경우 해당 작업에서 낮은 성능을 보여주는 현상에 기인한 것으로 보인다. 미세 조정된 BERT 기반의 키워드 추출 모델들은 모두 약 0.28 이상의 F1-Score를 보여주었다. 그중 생의학 분야의 데이터로 사전학습 된 BioBERT의 F1-Score가 0.315로 가장 높은 성능을 보여주었다. 이를 통해 생의학 분야의 키워드 추출 진행 시 미세 조정한 BioBERT 모델을 사용하는 것이 가장 효과적이라는 사실을 발견했다.

4.2 키워드 추출 결과 비교

실험에 사용된 평가 데이터 중 실제 논문 초록과 키워드를 예시로 모델별 키워드 추출 결과를 비교하였으며 그 결

(표 5) 키워드 추출 모델별 성능 비교 결과
(Table 5) Performance comparison results by keyword extraction model

		P	R	F
비지도 학습 기반	FirstPhrases	.095	.118	.101
	TextRank	.032	.049	.037
	SingleRank	.062	.084	.068
	TopicRank	.125	.145	.129
	PositionRank	.073	.096	.079
	Multipartile Rank	.140	.165	.146
	YAKE	.136	.160	.141
BERT 기반 (Base)	BERT	.015	.056	.023
	DistillBERT	.016	.035	.021
	XL-NET	.001	.001	.001
	ALBERT	.026	.037	.030
	KeyBERT	.027	.021	.022
	BioBERT	.007	.017	.01
BERT 기반 (Fine-Tuned)	BERT	.309	.319	.303
	DistillBERT	.321	.264	.279
	XL-NET	.31	.327	.31
	ALBERT	**.324**	.269	.282
	BioBERT	.323	**.328**	**.315**

과는 표 6과 같다. 표 6의 예시 문장은 살충제와 감자의 질병 간의 관계에 관한 내용이며 키워드는 감자 난종 선충(potato cyst nematode), 선충(nematode), 살충제(nematicide)와 같은 단어들이다.

비지도 학습 기반의 키워드 추출 모델은 예시 키워드와 같은 개수를 확률값이 높은 순서대로 추출하였으며 BERT 기반 키워드 추출 모델은 키워드로 분류되는 단어들을 모두 추출하였다. 비지도 기반 키워드 추출 모델 중 가장 좋은 성능을 보여준 MultipartileRank 모델은 선충 살충제(cadusafos), 부화(hatching)와 같은 관련된 키워드를 추출하였으나 효과(effect), 침입(invasion)과 같은 관련 없는 단어도 추출하는 것을 확인할 수 있다. 반면 BERT 기반 키워드 추출 모델 중 가장 좋은 성능을 보여준 미세 조정된 BioBERT 모델은 선충 살충제(cadusafos), 부화(hatching)등 예시 키워드와 관련 있는 키워드만 추출한 것을 확인할 수 있다.

5. 결론 및 향후 연구

본 논문에서는 생의학 분야의 기법별 다양한 키워드 추출 모델을 사용하여 모델별 성능을 비교하고 추출된 키워드를 제시했다. BioBERT는 F1-Score 기준 0.315로 가장 높은 성능을 보여주었다. 이를 통해 생의학 분야에서 BERT의 구조나 환경을 변화시켜 성능을 증가시킨 모델 보다 해당 분야의 데이터로 학습시킨 모델을 사용하는 것이 키워드 추출에서 더욱 효과적이라는 것을 확인했다.

향후 연구에서는 임베딩 기반 키워드 추출 모델들을 포함한 비지도 학습 기반의 키워드 추출 모델 중 가장 좋은 성능을 보여주고 있는 모델인 PromptRank[12]를 추가하여 연구를 진행해 더욱 종합적인 비교를 제공할 수 있을 것으로 기대된다. 또한 생의학 분야뿐만 아니라 금융, 과학 등 다양한 분야의 키워드 추출 모델 비교 연구와 추출된 양질의 키워드를 활용한 후속 연구들이 진행될 수 있을 것으로 기대된다.

(표 6) 키워드 추출 모델별 키워드 추출 예시
(Table 6) Example keyword extraction by keyword extraction model

예시 문장	cadusafos inhibits hatching invasion, and movement of the potato cyst nematode globodera pallida. the effect of the nematicide cadusafos on the hatching of the potato cyst nematode globodera pallida in potato root diffusate, soil leachate, and distilled water was investigated. cadusafos had a significant effect on the hatching, migration, movement, and root invasion by the second-stage juveniles. hatching was completely inhibited at low concentrations of cadusafos 0.002-0.004 microg/ml, but hatching resumed a week after removing the nematicide. at concentrations of 0.05 microg/ml and higher of analytical-grade cadusafos, the inhibition of hatching was permanent.	
예시 키워드	['globodera pallida', 'cadusafos', 'hatching', 'nematicide', 'nematode', 'potato cyst nematode']	
	키워드 추출 모델	**추출 키워드**
비지도 학습 기반	FirstPhrases	cadusafos, invasion, movement, potato cyst nematode globodera pallida, effect
	TextRank	potato cyst nematode globodera pallida, root invasion, analytical-grade cadusafos, nematicide cadusafos, low concentrations
	SingleRank	potato cyst nematode globodera pallida, nematicide cadusafos, analytical-grade cadusafos, cadusafos, root invasion
	TopicRank	hatching, cadusafos, potato cyst nematode globodera pallida, movement, effect
	PositionRank	nematicide cadusafos, analytical-grade cadusafos, potato root diffusate, cadusafos, root invasion
	MultipartileRank	cadusafos, hatching, potato cyst nematode globodera pallida, effect, invasion
BERT 기반 (Base)	YAKE	potato cyst nematode, cyst nematode globodera, nematode globodera pallida, inhibits hatching invasion, hatching
	BERT	hatching was permanent, inhibits, investigated cad, juveniles hatching, of the potato cyst
	DistillBERT	and root invasion by, soil leach, the effect of, in potato root diffusa, inhibits
	XL-NET	X
	ALBERT	migration, movement, and root invasion by the second, soil leachate, at concentrations of 0
	KeyBERT	migration, juveniles, invastion, root, low
	BioBERT	and, concentrations, ne, on
BERT 기반 (Fine-Tuned)	BERT	cadusafos, hatching, nematic, potato, potato cyst nematode globodera pallida
	DistillBERT	cadusafos, hatch, nematic, potato, potato cyst nematode globodera pallida
	XL-NET	cadusafos, hatching, nematicide cadusafos, nematode globodera pallida, potato
	ALBERT	cadusafos, hatching, matic, movement, potato cyst nematode globodera pallida
	BioBERT	cadusafos, hatch, hatching, nematic, potato cyst nematode globodera pallida

참고문헌(Reference)

[1] Beliga, Slobodan, Ana Meštrović, and Sanda Martinčić-Ipšić, “An overview of graph-based keyword extraction methods and approaches,” Journal of information and organizational sciences, Vol 39, No.1, pp.01-20, 2015.
https://hrcak.srce.hr/140857

[2] Gero, Zelalem, and Joyce C. Ho, “Namedkeys: Unsupervised keyphrase extraction for biomedical documents,” Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics, pp.328-337, 2019.
https://doi.org/10.1145/3307339.3342147

[3] A. Çelikten, A. Uğur and H. Bulut, “Keyword Extraction from Biomedical Documents Using Deep Contextualized Embeddings,” 2021 International Conference on INnovations in Intelligent SysTems and Applications (INISTA), Kocaeli, Turkey, 2021.
https://doi.org/10.1109/INISTA52262.2021.9548470

[4] Campos, Ricardo, et al., “YAKE! Keyword extraction from single documents using multiple local features,” Information Sciences 509, pp.257-289l, 2020.
https://doi.org/10.1016/j.ins.2019.09.013

[5] Mihalcea, Rada, and Paul Tarau. “Textrank: Bringing order into text,” Proceedings of the 2004 conference on empirical methods in natural language processing. 2004.
https://aclanthology.org/W04-3252

[6] Brin, Sergey, and Lawrence Page, “The anatomy of a large-scale hypertextual web search engine,” Computer networks and ISDN systems, Vol 30, No.1-7, pp107-117, 1998.
https://doi.org/10.1016/S0169-7552(98)00110-X

[7] Wan, Xiaojun, and Jianguo Xiao, “Single document keyphrase extraction using neighborhood knowledge,” AAAI. Vol. 8. 2008.
https://cdn.aaai.org/AAAI/2008/AAAI08-136.pdf

[8] Bougouin, Adrien, Florian Boudin, and Beatrice Daille, “Topicrank: Graph-based topic ranking for keyphrase extraction,” International joint conference on natural language processing (IJCNLP), 2013.
https://aclanthology.org/I13-1062/

[9] Bennani-Smires, Kamil et al., “Simple unsupervised keyphrase extraction using sentence embeddings,” arXiv preprint arXiv:1801.04470 2018.
https://arxiv.org/abs/1801.04470

[10] Boudin, Florian, “Unsupervised keyphrase extraction with multipartite graphs,” arXiv preprint arXiv:1803.08721 2018.
https://arxiv.org/abs/1803.08721

[11] Y. Sun, H. Qiu, Y. Zheng, Z. Wang and C. Zhang, “SIFRank: A New Baseline for Unsupervised Keyphrase Extraction Based on Pre-Trained Language Model,” in IEEE Access, vol. 8, pp. 10896-10906, 2020.
https://ieeexplore.ieee.org/abstract/document/8954611

[12] Kong, Aobo, et al. ″PromptRank: Unsupervised Keyphrase Extraction Using Prompt.″ arXiv preprint arXiv:2305.04490 2023.
https://doi.org/10.48550/arXiv.2305.0449

[13] Devlin, Jacob, et al. ″Bert: Pre-training of deep bidirectional transformers for language understanding.″ arXiv preprint arXiv:1810.04805 2018.
https://doi.org/10.48550/arXiv.1810.04805

[14] Liu, Yinhan, et al., “Roberta: A robustly optimized bert pretraining approach,” arXiv preprint arXiv:1907.11692 2019.
https://doi.org/10.48550/arXiv.1907.11692

[15] Lan, Zhenzhong, et al., “Albert: A lite bert for self-supervised learning of language representations,”
https://doi.org/10.48550/arXiv.1909.11942

[16] Lee, Jinhyuk, et al., “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” Bioinformatics Vol.36. No.4, pp.1234-1240. 2020.
https://doi.org/10.1093/bioinformatics/btz682

[17] Beltagy, Iz, Kyle Lo, and Arman Cohan, “SciBERT: A pretrained language model for scientific text,” arXiv preprint arXiv:1903.10676 2019.
https://doi.org/10.48550/arXiv.1903.10676

[18] Ramshaw, Lance A., and Mitchell P. Marcus. “Text chunking using transformation-based learning,” Natural language processing using very large corpora. Dordrecht: Springer Netherlands, pp.157-176, 1999.
https://doi.org/10.1007/978-94-017-2390-9_10

[19] Florescu, Corina, and Cornelia Caragea, “Positionrank: An unsupervised approach to keyphrase extraction from

scholarly documents," Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: long papers). 2017.
https://aclanthology.org/P17-1102/

[20] Sanh, Victor, et al., "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," arXiv preprint arXiv:1910.01108, 2019.
https://doi.org/10.48550/arXiv.1910.01108

[21] Yang, Zhilin, et al., "Xlnet: Generalized autoregressive pretraining for language understanding," Advances in neural information processing systems 32 2019.
https://doi.org/10.48550/arXiv.1907.11692

◐ 저 자 소 개 ◑

이 동 희(Dong-hee Lee)
2022년 동국대학교 경영학과(경영학사)
2023년~현재 연세대학교 정보대학원
비즈니스 빅데이터 분석 트랙 석사과정
관심분야 : Natural Language Processing, Deep Learning
E-mail : dlehdgml1031@yonsei.ac.kr

권 순 찬(Soon-chan Kwon)
2022년 광운대학교 행정학 학사
2023년~현재 연세대학교 정보대학원
비즈니스 빅데이터 분석 트랙 석사과정
관심분야 : Natural Language Processing, Deep Learning
E-mail : elrnjs13@yonsei.ac.kr

장 백 철(Beakcheol Jang)
2009년 North Carolina State University 컴퓨터공학과(공학박사)
2021년~현재 연세대학교 정보대학원 교수
관심분야 : Natural Language Processing, Bigdata Analysis
E-mail : bjang@yonsei.ac.kr