연구 동향 분석을 위한 텍스트 마이닝 기반 GPT 활용 기법

하정훈[O], 최봉준[*]
[O]동서대학교 소프트웨어학과,
[*]동서대학교 소프트웨어학과
e-mail: logo8044@gmail.com[O], bongjun.choi@dongseo.ac.kr[*]

Text mining based GPT utilization technique for research trend analysis

Jeong-Hoon Ha[O], Bong-Jun Choi[*]
[O]Dept. of Software, Dongseo University,
[*]Dept. of Software, Dongseo University

● 요 약 ●

새로운 연구를 시작하기 위해서는 과거의 연구 동향을 분석해야 한다. 이를 위해 많은 양의 과거 연구 데이터를 조사해야 하는데, 모든 데이터를 직접 분류하는 방법은 많은 시간과 노력이 필요하기 때문에 비효율적이며, 텍스트 마이닝 기법을 활용한 키워드분석만으로는 연구 동향을 이해하기에 어려움이 존재한다. 이러한 전통적인 키워드 추출 방법의 한계점을 보완하기 위해 본 논문에서는 텍스트 마이닝 기반 GPT 활용 기법을 제안한다. 본 연구에서는 특정 도메인에 대해 텍스트 마이닝 기법을 활용하여 키워드를 추출하고, 이러한 키워드를 해당 도메인의 데이터로 미세 조정(fine-tuning)된 GPT의 입력으로 사용한다. GPT 결과로 생성된 문장을 텍스트 마이닝으로 나온 결과와 비교 분석한다. 이를 통해 연구 분야의 동향 분석을 보다 쉽게 할 수 있을 것으로 기대된다.

키워드: GPT(Generative Pre-trained Transformer), 텍스트 마이닝(text mining), 연구 동향(research trends)

I. Introduction

기존의 텍스트 마이닝은 자연어로 구성된 텍스트 데이터를 대상으로 의미있는 정보를 생산하기 위해 관계 패턴을 도출하여 대용량의 비정형 텍스트 데이터로부터 동향을 분석하는데 많이 활용되고 있다. 텍스트 마이닝을 활용하여 학술적으로 의미 있는 텍스트로 구성된 빅데이터라는 점을 가진 논문의 연구 동향을 분석하는 것은 여러 분야에서 수행되어 왔으며 의미 있는 연구결과를 도출하고 있다. [1]

그러나 모든 논문을 읽고 연구 내용을 정리하여 연구 동향을 분석하는 작업은 현실적으로 불가능하다. 텍스트 마이닝을 통해 추출된 키워드만을 가지고 연구 동향을 분석하고 이해하기에 한계가 존재하기 때문에 키워드기법을 통한 분석에 개선이 필요하다.

GPT는 다양한 문맥과 문장을 이해하고 생성할 수 있는 강력한 자연어 처리 모델이지만, 그 자체로는 텍스트 데이터와 관련된 문장을 생성하지 못하는 경우가 다분하다.

본 논문에서는 이러한 한계점을 극복하기 위해 텍스트 마이닝의 결과인 키워드를 허깅페이스(HuggingFace)의 GPT를 기반으로 미세조정(fine-tuning)한 모델의 입력값으로 넣어 생성된 문장을 비교 분석한 연구 내용을 기술한다.

II. Preliminaries

텍스트 마이닝 기법은 텍스트로부터 지식을 발견하고 추출하는 것이며, 텍스트 마이닝 기법을 도입하면 객관적인 분석이 가능하다. 본 연구에서는 텍스트 마이닝을 논문의 연구 동향에 적용하여 분석하는 기법으로 불용어를 처리하고 빈도 분석을 통한 스코어링을 사용하여 키워드를 추출하는 방법을 사용한다.

GPT는 OpenAI에서 개발한 대규모 언어 모델로, 15억개의 파라미터를 가지며 zero-shot 환경에서 8개 중 7개에서 state-of-the-art를 달성하였다. 최종 결과는 다시 단어로 변환 및 조합되어 출력한다. [2]

허깅페이스(HuggingFace)는 다양한 트랜스포머 모델(transformer models)과 학습 스크립트(transformer Trainer)를 제공

하는 모듈이다. 허깅페이스(HuggingFace)에서 제공한 GPT 모델은 사용시 layer, model 등을 선언하거나 학습 스크립트를 구현하지 않아도 된다는 장점이 있다. [3]

기존에 OpenAI에서 제공하는 GPT는 프롬프트를 지정하여 그 뒤에 문장에 생성되는 구조이기 때문에 파라미터 값을 조정하여 미세 조정(fine-tuning)해도 텍스트 마이닝 기법을 이용해 추출해낸 키워드 뒤에 어색한 문장이 생성되게 된다.

본 연구에 사용된 허깅페이스(HuggingFace)에서 제공하는 GPT 모델은 관련 키워드가 자연스러운 위치에 들어가 문장이 생성되기 때문에 OpenAI에서 제공하는 모델보다 더 자연스러운 문장이 만들어진다.

III. The Proposed Scheme

1. System Architecture

본 논문에서 제안하고자 하는 시스템은 다음과 같은 구조로 구성되어 있다. 사용자가 분석할 텍스트 파일을 선택하고 추출하고자 하는 상위 키워드 개수를 입력하면, 그에 맞는 키워드와 스코어가 추출된다. 문장 생성 버튼을 누르면 추출된 키워드를 기반으로 전처리된 데이터가 입력된 GPT를 통해 연산을 수행하여 문장이 생성된다. 생성된 문장은 데이터베이스에 저장되거나 필요에 따라 출력될 수 있다.

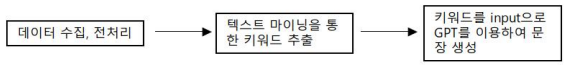

Fig. 1. system concept

이러한 시스템 구조를 통해 방대한 양의 텍스트 데이터에서 상위 키워드를 추출하고, 그에 맞는 문장을 생성하여 동향을 파악할 수 있다.

2. Data Analysis

본 연구에서 데이터는 특정 주제 분야 연구를 위해 scopus api를 활용하여 수집하였다. 이를 기반으로 GPT의 사전 학습과 미세 조정(fine-tuning)을 하였다. 해당 데이터는 연도 별로 수집된 논문들로 이루어져 있다.

Table 1. FIne-tuning settings

Max_length	50	temperature	0.8
Epochs	1	Sequences	1
Batch_size	8	Blcok_size	1024

3. System Result

텍스트 마이닝한 결과와 이를 통해 생성된 문장은 다음과 같다.

Table 2. System results

vehicle	This paper addresses the planning issue of empty vehicle management.
control	Logistics control issues and quantitative decision support.

IV. Conclusions

본 논문에서는 텍스트 마이닝 기법을 사용하여 사용자가 분석하고자 하는 텍스트 데이터 파일에서 상위 키워드를 추출하고 이 키워드를 GPT의 입력으로 넣어 생성되는 결과인 문장을 비교 분석하는 시스템의 개발 내용에 관하여 기술하였다.

본 시스템을 통해 사용자는 연구하고자 하는 분야의 동향 분석을 보다 손쉽게 할 수 있으며, 새로운 연구를 시작하기 전에 연구 동향 분석을 할 때 연구 주제와 상관없이 다양한 분야에서 활용될 수 있을 것으로 기대된다.

ACKNOWLEDGEMENT

본 연구는 2023년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음 (2019-0-01817)

REFERENCES

[1] Jaewoo Kim, Dong-jin Kim. “A Study on the Research Trends of Social Studies Using Text Mining : Focused on Academic Papers After 2000” Journal of Information Processing (2019): 35-70

[2] Hugging Face, https://huggingFace.co/docs/transformers/model_doc/gpt2

[3] Hugging Face, https://huggingface.co/docs/transformers/index