국가R&D과제정보 요약을 위한 한국어 정보요약 시스템

이종원 · 김태현 · 신동구 · 조우승*

한국과학기술정보연구원

Korean Information Summary System for National R&D Projcet Information Summary

Jong-Won-Lee · Tae-Hyun-Kim · Dong-Gu-Shin · Woo-Seung Jo*

Korea Institute of Science and Technology Information

E-mail : jwon1991@kisti.re.kr / heemang@kisti.re.kr / lovesin@kisti.re.kr / champion29@kisti.re.kr

요 약

국가과학기술지식정보서비스(이하 NTIS)에서는 국가R&D과제정보를 제공하고 있다. 과제정보는 '과제명', '과제수행기관', '연구책임자명' 등의 메타정보와 '연구목표', '연구내용', '기대효과'와 같은 과제를 설명하는 텍스트들로 구성되어있다. 과제정보 100만건을 대상으로 검색한 결과목록에서 '연구목표'나 '연구내용' 등을 모두 확인하여 원하는 과제정보를 찾기 위해서는 많은 시간이 필요하다는 문제가 있다. 이러한 문제점을 해소하기 위해, 본 논문에서는 국가R&D 과제정보 내에서 장문의 텍스트로 구성된 부분을 요약하는 과제정보 요약 시스템을 제안하고자 한다. 한국어의 언어학적 특징을 분석하여 전처리기를 구축하고 전처리된 텍스트 정보를 처리하기 위한 자연어 처리 기술 기반 과제정보 요약 모델을 개발하였다. 이를 통해 장문으로 구성된 과제정보를 압축 및 요약된 형태로 제공하여, 이용자들이 요약정보만으로도 전반적인 내용을 쉽고 빠르게 유추하는 데 도움이 될 것이다.

ABSTRACT

The National Science and Technology Knowledge Information Service (NTIS) provides information on national R&D projects. Project information consists of meta-information such as 'project name', 'project performance institution', 'research manager name', and text explaining projects such as 'research goal', 'research content', and 'expected effect'. There is a problem that it takes a lot of time to find the desired project information by checking all of the "research goals" or "research contents" in the list of results of searching for 1 million project information. To solve this problem, this paper proposes a project information summary system that summarizes the parts consisting of long texts within the national R&D project information. By analyzing the linguistic characteristics of the Korean language, a preprocessor was built and a project information summary model based on natural language processing technology was developed to process preprocessed text information. Through this, project information composed of long sentences is provided in a compressed and summarized form, which will help users to easily and quickly infer the overall content with the summary information alone.

키워드

Artificial Intelligence, linguistic Features, Natural Language Processing, NTIS, Project Information Summary

Ⅰ. 서 론

국가과학기술지식정보서비스(이하 NTIS)에서는 약 99만건 이상의 국가R&D과제정보를 제공하고 있다. 과제정보는 '과제수행기관', '과제명', "연구책임자명", '연구관리전문기관' 등의 메타정보와

* corresponding author

‘연구목표’, ‘연구내용’, ‘기대효과’ 와 같이 장문의 텍스트들로 구성된 정보가 있다. 연구자들은 장문의 텍스트를 읽고 원하는 과제를 찾아내는데 많은 시간과 노력이 요구된다. 또한, NTIS를 활용하는 연구자는 국가R&D과제정보를 가장 많이 활용하기 때문에 NTIS 이용자들에게 효율적으로 정보를 제공하기 위해서는 국가R&D과제정보를 압축하여 중요한 정보를 제공해주는 기술이나 시스템이 필요한 실정이다[1].

이를 위해 본 논문에서는 국가R&D과제정보 중 장문의 텍스트로 구성된 부분을 요약하는 시스템을 제안한다.

본 시스템은 한국어의 언어학적 특징을 분석하여 텍스트를 처리하는 전처리기와 자연어 처리 기술 기반 한국어 요약 모델로 구성된다. 제안 시스템을 활용함으로써 이용자들이 장문으로 구성된 국가R&D과제정보를 압축 및 요약된 형태로 우선 확인할 수 있어, 정보 획득이 보다 용이해질 것으로 기대된다[2-4].

II. 본 론

제안하는 시스템은 전처리기와 자연어 처리 기술 기반 인공지능 모델로 구성되어있다. 그림 1은 시스템의 흐름도이다.

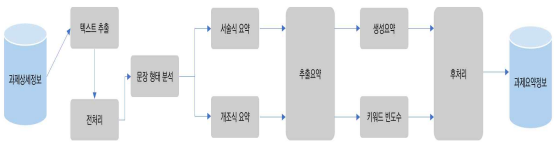

그림 1. 시스템 흐름도

시스템은 우선 국가R&D과제정보에서 텍스트를 추출한다. ‘연구목표’와 ‘연구내용’을 활용하며 ‘기대효과’는 과제의 연구내용을 대표하지 않는 문맥적 흐름을 갖기 때문에 활용하지 않는다. 전처리부에서는 불용어 처리, 띄어쓰기 확인, 특수문자 제거, 특정 문구 출현 확인(본 연구는, 본 논문은, 본 과제는 등)을 수행한다. 한국어의 언어학적 특징으로 인해 ‘본 연구는’, ‘본 과제는’ 등의 문구 뒤에는 핵심 문장이 올 확률이 높으므로 이를 활용하기 위한 것이다. 그리고 ‘연구목표’와 ‘연구내용’의 문장 형태가 서술식인지 개조식인지를 확인한다. 일반적인 자연어 처리 기술 기반 인공지능 모델은 개조식 문장을 잘 이해하지 못한다. 그 이유는 문장의 끝맺음 부분이 서술어가 아니기 때문이다. 따라서 서술식 문장 형태의 과제정보인 경우 추출요약과 생성요약을 같이 활용하여 요약문을 생성하고, 개조식 문장 형태의 과제정보인 경우 추출요약과 키워드 빈도수 기반으로 핵심 문장을 선별하는 방식으로 요약문장을 추출한다. 추출요약 모델은 입력된 텍스트 정보에서 중요도가 높은 3개의 문장을 도출하는 역할을 수행하며 생성요약 모델은 추출요약 모델이 도출한 문장을 1개의 문장으로 압축 및 요약하는 역할을 수행한다. 키워드 빈도수 기반 핵심 문장 선별은 TF-IDF 기법을 활용하며, 후처리부에서는 단어가 깨지는 현상이 발생하는 경우 이를 해결하기 위해 국가R&D과제정보에 포함되어있는 키워드 정보를 활용한다. 해당 키워드는 국가R&D과제의 책임자가 과제정보를 입력할 때 과제를 대표하는 용어로써 기입한 것이기에 단어보정을 위해 활용하기에 적합하다.

제안한 시스템을 적용하여 그림 2와 같은 R&D 정보요약 서비스를 구축하였다.

그림 2. R&D정보요약 서비스 화면

본 R&D정보요약 서비스는 ‘연구목표’와 ‘연구내용’을 압축 및 요약하여 이용자들에게 국가R&D과제정보에 대한 핵심정보를 제공해주고 있다. ‘연구목표’는 국가R&D과제의 목표정보를 담고 있기 때문에 의미적으로 요약 문장의 주어 역할에 해당된다고 볼 수 있고, ‘연구내용’은 과제에 대한 구체적인 연구 내용을 담고 있어 의미적으로 요약정보 내에서 주어를 뒷받침하는 서술어 부분에 해당된다고 볼 수 있다. 이러한 언어학적 특징을 적용하여 시스템이 도출한 과제정보 요약은 국가R&D과제를 수행한 기관이나 기업이 ‘연구목표를 위해 연구내용을 수행하였다’라는 구조로 요약정보를 생성할 수 있게 함으로써 보다 유의미한 요약정보를 이용자들에게 제공할 수 있다.

III. 결 론

장문의 텍스트로 구성된 정보를 분석하고 활용하기 위해서는 많은 시간과 노력이 요구된다. NTIS에서 제공하고 있는 국가R&D과제정보 내에도 '연구목표', '연구내용'과 같이 장문의 텍스트 정보가 많으므로 이용자들이 해당 정보를 분석하고 활용하기 위해서 많은 시간과 노력을 쏟고 있다. 본 연구에서는 이러한 점을 개선하기 위해 국가R&D과제정보를 압축 및 요약하고 중요한 정보를 제공해주는 시스템을 구축하였다. 본 시스템은 한국어의 언어학적 특징을 분석하고 처리하는 전처리기와 요약정보 생성을 위한 자연어 처리 기술 기반 한국어 요약모델을 적용하였다.

제안하는 시스템은 장문의 텍스트로 구성된 국가R&D과제정보를 분석하고 압축 및 요약하여 유의미한 내용으로 구성된 단문의 텍스트 정보를 이용자에게 제공할 수 있게 한다. 이를 통해 NTIS 이용자들이 국가R&D과제정보를 보다 쉽게 활용하여 연구자들의 연구 생산성을 향상할 수 있을 것으로 기대한다.

Acknowledgement

이 논문은 2022년도 한국과학기술정보연구원(KISTI) 주요사업 과제(NTIS 과제고유번호 1711173845) 지원에 의함.

References

[1] K. H. Lee, S. H. Na, J. H. Lim, T. H. Kim, and D. S. Chang, "PrefixLM for Korean Text Summarization," *Journal of Korean Institute of Information Scientists and Engineers*, Vol. 49, No. 6, pp. 475-487, Jun. 2022.

[2] J. H. Shin, E. H. Kim, and M. J. Lim, "Improving the effectiveness of document extraction summary based on the amount of sentence information," *Smart media journal*, Vol. 11, No. 3, pp. 31-38, Apr. 2022.

[3] J. H. Cho, and H. Y. Oh, "Training Techniques for Data Bias Problem on Deep Learning Text Summarization," *Journal of the Korea Institute of Information and Communication Engineering*, Vol. 26, No. 7, pp. 949-955, Jul. 2022.

[4] G. H. Lee, Y. H. Park, and K. J. Lee, "Deletion-based Korean Sentence Compression using Graph Neural Networks," *Journal of Korean Institute of Information Scientists and Engineers*, Vol. 49, No. 1, pp. 32-41, Jan. 2022.