논문 구조 기반 문장 범주를 이용한 AI 논문 요약 서비스

AI paper summarization service using sentence classification based on the structure of academic paper

선 충 녕, 이 태 석, 강 남 규, 이 석 형
한국과학기술정보연구원, 융합서비스센터

Seon Choong-nyoung, Lee Tae-Seok,
Kang Nam Gyu, Lee Seok Hyoung
Korea Institute of Science and Technology Information, Convergence Service Center

요약

과학기술에 대한 논문 출판의 수는 빠른 속도로 증가되고 있으며, 이런 상황에서 연구자들은 자신들이 원하는 논문에 접근하는데 어려움을 겪는다. 논문 접근의 수단으로 사용되던 메타 정보는 연구자들을 효과적으로 돕는 데 한계가 있어 다양한 방법들이 제안되었다. 본 연구에서는 논문 구조에 기반해 원문 문장을 분류하는 기법을 소개한다. 이를 이용하여 다양한 관점에서 접근하기를 원하는 연구자들을 지원하는 AI 논문 요약 서비스를 구축하였다.

I. 서론

최근 과학기술에 대한 학술논문의 수가 폭발적으로 증가하고 있다. 연구자들은 논문을 읽고 이해하는데 많은 시간이 필요하므로 찾은 논문이 자신이 원하는 논문인지를 확인할 방법이 필요하다. 전통적으로 저자 초록이 제공되고 있지만, 대부분의 초록은 구조화되지 않기 때문에 원하는 부분을 발췌하여 읽는데 어렵고, 논문 중에서 방법과 결론 외에 다른 부분에 관심이 있더라도 초록에 포함되지 않는 경우가 많아 연구자들이 다양한 관점에서 원하는 논문을 정확히 판별하는 데 한계가 있었다.

이런 상황에서 논문의 원문을 보지 않고도 원하는 논문인지 확인하기 위한 다양한 방법이 제안되고 있다. Cachola는 검색 결과 목록에서 짧은 문장으로 논문을 표현해주는 TLDR(too long; didn't read)이라는 이름의 극단적으로 짧은 요약을 제공하는 연구를 진행하였다[1]. Semantic Scholar에 직접 적용된 이 연구는 초록보다 짧게 논문의 요지를 파악할 수 있게 해주기 때문에 논문을 선별하는 데 도움을 줄 수 있다. 송민선은 의미적인 메타 정보를 구축하여 사용자의 관점에 맞는 정보에 손쉽게 접근할 수 있도록 의도를 이용한 검색을 할 수 있는 연구를 진행하였다[2].

하지만 Cachoia의 방법은 극단적으로 짧은 요약문만을 제공하므로 다양한 관점에 대한 접근이 필요한 경우 초록과 원문에 접근이 필요하며, 의미적인 메타 정보의 경우는 별도로 구축을 해주어야 하므로 수많은 논문에 적용하는 데 한계가 있었다.

우리는 이러한 선행연구들의 한계를 극복하려는 방법을 제안한다. 사용자들의 다양한 관점에 대응할 수 있도록 논문의 구조적 역할에 맞는 문장 태그를 정의하였다. 이렇게 정의된 태그를 논문 원문에 적용하여 각 역할을 대표할 수 있는 문장들을 추출하고, 이를 통해 논문을 이해하기 위한 요약 정보를 활용하였다.

II. 구조정보를 이용한 논문 문장 의미 태그

제안된 방법은 논문의 원문에서 중요한 역할을 하는 문장들을 선별하여 다양한 관점에서 내용에 접근할 수 있는 논문의 대표적인 정보를 구축하여 활용한다. 이를 위해서 논문을 연구주제, 연구방법, 연구결과의 관점에서 접근하였다.

보다 구체적으로 정의하기 위해 각각의 관점은 세부적인 범주로 상세화하였다. 연구주제는 문제 정의, 가설설정, 기술 정의로 구분하였고, 연구방법은 제안 방법, 이론&모형, 대상 데이터, 데이터처리로 세분화하였다. 마지막으로 연구결과는 성능/효과, 후속연구로 구분하였다[3].

수많은 논문에 대해 각 범주를 대표하는 문장을 수동 구축하는 것은 불가능하므로 기계학습 방법을 통해 원문에서 추출하고자 하였다. 학습용 데이터 구축은 대학교 졸업 이상의 학력을 가진 사람들이 진행하였으며 14,086개의 논문에 대해 155,740문장의 논문 문장 의미 태그 학습데이터를 구축하였다[4].

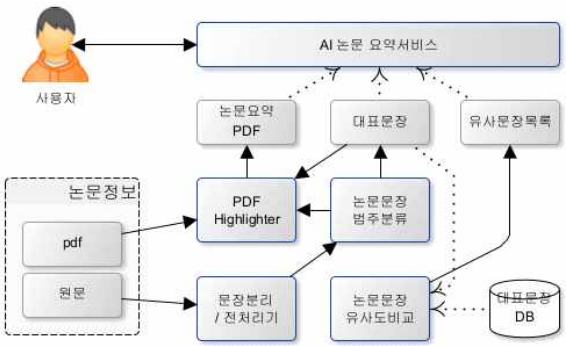
III. AI 논문 요약서비스 설계 및 구현

1. AI 논문 요약서비스 설계

구축된 데이터를 이용하여 역할별 대표문장을 추출할 수 있는 모델을 개발하였고 이를 통해 사용자들이 논문에 효율적으로 접근할 수 있는 AI 논문 요약서비스를 설계하였다.

AI 논문 요약서비스는 논문을 검색할 때 연구주제, 연구방법, 연구결과로 구분된 문장을 확인할 수 있으며, 관심이 있는 논문의 경우 상세보기에서 각 의미 범주별 대표문장들을 제공하여 원하는 논문이 맞는지 원문을 보지 않고도 확인할 수 있도록 설계하였다.

추가로, 유사 문장 검색 모듈을 이용하여 비슷한 의미의 문장을 가진 논문을 찾을 수 있는 기능과 원하는 논문의 경우 범주로 선택된 문장들을 표시하여 읽으면서 도움을 받을 수 있는 논문 요약 PDF 기능을 구성하였다. 그림 1은 이와 같은 논문 요약서비스의 개요를 표현하고 있다.

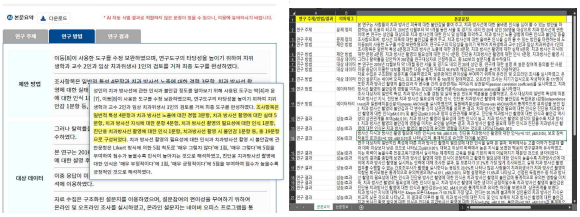


▶▶ 그림 1. AI 논문 요약 서비스 개요

2. AI 논문 요약 서비스 구현

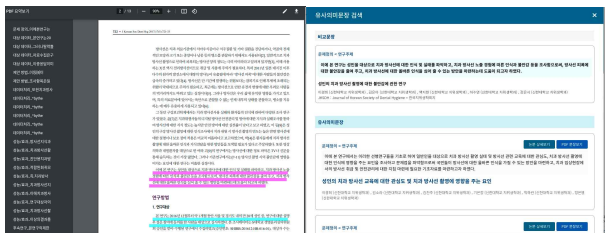
AI 논문 요약 서비스는 ScienceON에 적용된 기능으로 크게 두 가지 방법으로 이용할 수 있다.

ScienceON 사용자는 통합검색을 통해 AI 논문 요약이 적용된 논문에 접근할 수 있다. 이 경우 AI 논문 요약 정보를 상세보기에서 제공하여 사용자에게 논문의 내용을 파악하기 쉽게 돕는다. AI 논문 요약이 적용된 논문의 상세보기에서는 논문의 역할에 따른 대표문장을 보고, 해당 문장이 출현한 구절을 확인할 수 있으며, 추출된 문장들과 논문 정보를 별도의 파일로 받을 수 있는 기능을 제공한다.



▶▶ 그림 2. 논문 상세보기 내 요약 정보 및 다운로드 결과

또한, 사용자는 베타 서비스를 통해 AI 논문 요약이 적용된 논문들을 대상으로 다양한 기능을 활용할 수 있다. 베타 서비스에서는 목록에서 하이라이트 된 PDF를 받고 유사한 의미의 문장이 포함된 다른 논문을 검색하는 등의 베타 기능도 포함되어 이용해 볼 수 있다.



▶▶ 그림 3. 베타 서비스 특화기능 (논문 요약 PDF, 유사논문검색)

Ⅳ. 결론 및 향후 과제

본 연구에서는 과학기술 분야 연구논문 출판 수량의 증가에 따라 연구자들을 지원해주기 위한 논문의 구조 기반 정보 추출 방법을 제안하였다. 선행연구들이 논문의 중심 내용 위주로 정리하거나, 수동으로 구축된 정보에 의존했던 것에 비해, 제안 방법은 다양한 관점을 가지는 사용자들이 초록이나 원문을 읽는 노력 없이 논문의 내용을 파악할 수 있도록 설계하였으며, 다양한 기능을 구현하여 사용자가 이용해 볼 수 있도록 ScienceON 서비스에 적용하였다.

다만, 현재의 서비스는 베타 서비스로 신뢰할 수 있는 결과가 확보된 논문을 대상으로 서비스가 제공되므로 이용이 제한적이다. 사용자에게 실질적인 도움을 제공하기 위해 모델의 성능을 향상하고, 피드백을 받을 수 있는 체계를 보완하여 원문이 확보된 논문 대부분을 대상으로 서비스를 확장 적용이 필요하다. 또한, 이용자 피드백을 통해 의견을 수렴한 결과 해외 SCI 논문을 대상으로 확장되면 더 유용할 것이라는 사용자 의견을 확보할 수 있었다. 이를 위해 향후 다국어를 지원할 수 있는 연구가 진행될 필요가 있다.

▮ 참 고 문 헌 ▮

[1] Cachola, Isabel, Kyle Lo, Arman Cohan and Daniel S. Weld. “TLDR: Extreme Summarization of Scientific Documents.” FINDINGS (2020).

[2] 송민선, 고영만. (2015). “한국학 연구 논문의 의미 구조 기반 메타데이터 연구.” 한국도서관·정보학회지, 46(3), 277-299.

[3] 현미환, 선충녕 (2021). “과학기술 학술논문의 의미구조 기반 문장태깅 방법 연구.” 한국인터넷정보학회 추계학술대회논문집, 21(2), 267-268.

[4] 한국과학기술정보연구원 (2021). “국내 논문 문장 의미 태깅 데이터셋.” Version 1.0. 한국과학기술정보연구원. https://doi.org/10.23057/36.