BART 모델 기반의 긴 특허 문서 요약 시스템 개발

구동준○ 원경재 이아라 이동훈 박종민 심민기 김천구 김일곤*

경북대학교

dongjunkoo02@gmail.com, yui06031@gmail.com, eara0367@gmail.com, hy05205@gmail.com,
pjmin0101@gmail.com, minkisim@naver.com, droneprobe@naver.com, ikkim@knu.ac.kr

Development of Long Patent Summarization(PatSumm) Assistant System based on BART model

Dongjun Koo○ Kyung Jae Won Ah Ra Lee Donghun Lee
Jongmin Park Mingi Sim Cheongu Kim Il Kon Kim

Kyungpook National University

요 약

특허 문서는 발명에 대한 방대한 양의 정보를 담고 있어 기업 내 전략 수립 등 활용에 대한 수요가 높다. 하지만 문서 특성상 법률 및 기술 용어가 많고 길이가 길어 한 문서의 내용을 파악하는 데에는 상당히 오랜 시간이 소요된다. 따라서 본 연구는 기존 특허 핵심 내용 파악 과정에서 발생하는 문제점을 해결하기 위해 추출 및 생성 요약 기술을 활용한 특허 문서 요약 보조 시스템을 제안한다. 본 연구에서 제안하는 특허 문서 요약 보조 시스템은 긴 원문에서 핵심 정보를 도출하여 요약문을 생성함으로써 발명의 주요 내용을 빠르게 파악할 수 있게 한다. 또한 특허청 요약서 작성 가이드에 기반한 섹션별 요약문을 제공함으로써 특허 문서 활용을 희망하는 일반인 및 관련 분야 종사자의 업무 효율 향상에도 기여할 수 있다.

1. 서 론

4차 산업혁명과 함께 다양한 산업 내 디지털 전환 현상이 심화되며 '리걸 테크(LegalTech)'에 대한 관심이 증가하고 있다. 리걸 테크는 법률(legal)과 기술(Technology)의 합성어로 2010년 전후 형성되어 전 세계적으로 빠르게 성장하고 있는 신산업분야이다[1]. 초기에는 판례 수집 및 분석과 같은 빅데이터 기술이 대부분이었으나 최근 자연어 처리 및 인공지능 기술을 활용해 다양한 분야로 확장되고 있다. 미국, 유럽 등지에서는 이미 의사결정에 리걸 테크 산업을 활용하고 있으며, 대표적인 예로 2012년 삼성-애플의 특허 소송에 활용된 블랙스톤 디스커버리의 법무 자료조사 대행 인공지능 시스템이 있다.

국내에서는 법령 및 규제 문제로 리걸 테크 시장이 활성화되지는 않았지만, 최근 윕스(WIPS ON), 워트인텔리전스(KEYWERT) 등 특허 관련 스타트업이 등장하고 있으며 미국 리걸 테크 시장 형성 초기와 유사하게 주로 다량의 문헌 분석 결과를 제공한다. 이는 단시간 내 전반적인 기술 동향 파악에는 유용하지만 특정 문서의 핵심 내용을 살펴보기에는 어려움이 있다. 특허 문서에는 발명과 관련된 방대한 양의 정보가 혼재되어 있으며 법률 및 기술 용어가 많아 한 문서의 내용을 파악하는 데는 상당히 오랜 시간이 소요된다. 현재 특허 관련 서비스 제공 기업들은 특허 문서 내 요약서를 활용해 문서의 핵심 정보를 제공하고 있다. 하지만 요약서는 작성하는 사람에 따라 양적 및 질적인 부분에서 차이가 크기 때문에 문서에 따라 요약서를 통해 특허의 핵심 내용을 잘 파악할 수 없는 경우도 발생한다.

따라서, 본 연구에서는 특허 문서 핵심 내용 파악 과정에서 발생하는 어려움을 해소하기 위해 자연어 처리 및 인공지능 기술을 활용한 특허 문서 요약 보조 시스템, Patent Summarization(PatSumm) assistant를 제안한다. PatSumm 어시스턴트는 한국 특허 문서 특성에 적합한 추출요약 및 생성요약 기술을 사용해 특허청에서 제공하는 요약서 작성 가이드라인에 따라 특허 문서에 대한 요약문을 생성해주는 시스템이다. 본 연구에서 제안하는 PatSumm 어시스턴트는 특허 문서에 대한 일반인의 활용 수요뿐 아니라 변리사의 요약서 작성 시간 단축 등 해당 분야 종사자의 업무 효율 또한 향상하고자 한다.

2. 관련 연구

텍스트 요약은 원문 내 핵심 정보를 한눈에 파악할 수 있게 정제하는 기술로, 요약 방법에 따라 추출 및 생성 요약으로 나눌 수 있다[2]. 추출 요약은 원문 내 문장에 중요도를 매겨 핵심 문장을 선택하는 반면, 생성 요약은 원문에 없는 새로운 텍스트로 요약문을 생성한다. 2020년 발표된 Bidirectional Encoder Representation from Transformer(BART)는 대표적인 생성 요약 기술로 기존 BERT 및 GPT의 장점을 결합한 모델이다[3]. 사전 학습 모델은 태스크 의존도가 높음에도 불구하고 BART는 여러 태스크에서 좋은 성능을 보였으며 텍스트 요약 포함 다양한 자연어 처리 분야에서 State-of-the-art(SOTA)를 달성했다.

현재 텍스트 요약 분야의 주요 연구 과제 중 하나는 긴 문서 요약(Long text summarization)이다[4]. 요약문은 원문의 내용을 잘 반영하면서도 중요한 정보만을 간추려야 하는데 원문의 길이가 길수록 다량의 텍스트를 분석해야 하므로 계산 복잡도가 증가한다. 또한 원문 내 핵심이 아닌 노이즈 데이터가 많이 포함된 경우 어떤 내용이 중요한지 가려내기 쉽지 않다. 선행 연구에서는 이러한 문제점을 해결하기 위해 추출 요약과 생성 요약 기술을 결합하거나 분할 정복 알고리즘을 적용하는 등 다양한 노력을 기울이고 있으나 아직 한국어 긴 문서 요약에 대한 연구는 많이 수행되지 않았다.

* 본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음(2021-0-01082). 이 연구는 2020년도 산업통상자원부 및 산업기술평가관리원(KEIT) 연구비 지원에 의한 연구임(20011061). 이 논문은 2018년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2018R1A6A1A03025109)

3. 연구 방법

3.1 탐색적 데이터 분석(Exploratory Data Analysis)

본 연구에서는 특허 문서 요약 보조 시스템 PatSumm 어시스턴트 개발을 위해 한국지능정보사회진흥원 인공지능 학습용 데이터 구축사업을 통해 수집된 AIHub 논문자료 요약 데이터 셋 내 특허 문서 데이터 셋을 사용한다. 해당 데이터 셋은 특허 명세서 전체 및 섹션별 생성 요약 데이터 약 15만 건을 포함하고 있다. 본 연구에서는 먼저 특허 문서의 주요 특성을 파악하기 위해 해당 데이터 셋에 대한 탐색적 데이터 분석(Exploratory Data Analysis, EDA)을 수행한다.

데이터 셋은 크게 원문 및 생성 요약 학습을 위해 사람이 작성한 요약문으로 구성되어있다. 원문은 특허 명세서 내 특허 청구범위를 제외한 나머지 부분으로 평균 18,295바이트이다. 원문 내 섹션 제목, 부호 등을 제외하고 완전한 문장 형태를 구성하는 데이터만 추출해 보면 한 개의 문서는 평균 167개의 문장 및 2,680개의 명사로 이루어져 있으며 문장의 구성 또한 일반적인 단문이 아니라 계속해서 이어지는 만연체가 주를 이룬다. 텍스트 요약에 많이 쓰이는 국내외 데이터 셋과 비교했을 때 비교적 긴 편에 속함을 알 수 있다. 단어의 출현 빈도를 분석해보면 원문 내 5번 이하 등장하는 단어가 전체 등장 빈도에서 차지하는 비율이 0.024%, 10번 이하는 0.045%를 차지하는 것을 보아 문서 내 희소 단어가 많음을 나타낸다. 요약문은 평균 372바이트, 2개의 문장 및 68개의 명사로 이루어져 있다. novel n-gram으로 요약문 내 원문에 없는 단어의 비율을 추출해보면 unigram에서 0.92%, bigram 7.30%, trigram 13.00%의 값을 보인다. 또한, 원문과 요약문에서 공통으로 나타난 bigram, trigram 상위 키워드를 도출해보면 '본 발명은 ...에 관한 것이다.', '상기와 같은 문제점을 해결하기 위해...' 등 문서 내 고정성을 띠는 어휘들의 결합체를 발견할 수 있다.

분석 결과를 종합해보면 특허 문서는 원문 길이가 상당히 길기 때문에 Long Text Summarization 기술 적용이 필요하다. 또한 법률 및 기술 용어를 포함하고 있는 전문적인 문서로 개략적인 내용보다는 정확한 정보 전달이 요구되기 때문에 요약문에도 원문과 통일된 용어 및 특허 문서에서 나타나는 정형화된 관용어구를 사용하였음을 알 수 있다.

3.2 Patent Summarization(PatSumm) Assistant 개발

본 연구에서는 EDA를 통해 도출한 특허 문서의 특성을 바탕으로 긴 특허 원문에 대한 요약문을 생성하는 PatSumm 어시스턴트를 개발한다. PatSumm 어시스턴트 개발에 사용된 데이터는 총 5만 건으로 전체 데이터 셋을 원문 길이에 따라 사분위로 나눈 후 각 범위에서 랜덤하게 동일한 수의 샘플을 추출하여 구성한다. PatSumm 어시스턴트 개발은 크게 가공, 추출 및 생성의 세 단계로 구성되어 있다. 가공 및 추출 단계에서는 긴 문서 요약 시 발생하는 계산복잡도 문제를 해결하기 위해 입력 데이터의 길이를 줄이는데, 이때 노이즈는 제거하면서 핵심 정보의 손실은 최소화하기 위해 요약문과의 Recall-Oriented Understudy for Gisting Evaluation(Rouge) Score를 이용한다.

먼저, 가공 단계에서는 특허청 요약서 작성 가이드라인을 기반으로 원문 내 요약서 생성에 필요한 핵심 정보를 포함하고 있는 하위 섹션을 도출하고 문장 분리 등 자연어 처리 기술 적용을 위한 작업을 수행한다. 본 연구에서는 표 1과 같이 원문을 구성하는 하위 섹션과 요약문 간 Rouge Score를 계산하고 결과가 가장 높은 4개 섹션(기술 분야, 해결하려는 과제, 과제의 해결 수단, 발명의 효과)을 추출 단계의 입력 데이터로 사용한다.

표 1. 원문 섹션-요약문 간 ROUGE Score

성능 / 섹션명	F1 Score		
	Rouge-1	Rouge-2	Rouge-l
기술분야	0.383368	0.280348	0.321051
배경기술	0.279857	0.099812	0.152809
해결하려는 과제	0.411411	0.244686	0.288521
과제의 해결수단	0.377441	0.310173	0.324623
발명의 효과	0.441642	0.282929	0.322593
발명을 실시하기 위한 내용	0.148174	0.099923	0.107357

추출 단계는 긴 원문을 생성 요약 모델의 입력 데이터로 사용할 수 있도록 일부 문장을 추출해 길이를 줄이는 작업을 수행하는데, 이때 통계적 방법을 적용하여 문서의 핵심 내용이 포함되도록 중요한 문장을 추출하는 작업을 수행한다. 먼저 문장 및 단어의 중요도를 계산하기 위해 형태소 분석을 수행하는데, 한국어는 영어와 달리 어근과 접사에 의해 단어의 기능이 결정되는 교착어이기 때문에 형태소 분석 과정이 매우 복잡하다. 본 연구에서는 PatSumm 어시스턴트 개발에 사용할 형태소 분석기를 선정하기 위해 표 2와 같이 다양한 한국어 형태소 분석기의 특성 및 성능을 비교 분석한다. 특허 문서 5만 건을 대상으로 하나의 문서를 분석하는데 걸리는 평균 시간을 측정한 결과 Mecab는 적당한 수의 품사 태그를 가지면서도 타 형태소 분석기 대비 압도적인 결과를 보임을 알 수 있다.

추출 요약 모델 선정을 위해서는 문장 간 유사도를 이용해 핵심 정보를 추출하는 세 가지 알고리즘의 성능을 비교한다. 본 연구에서는 여러 선행 연구에서 기준 모델로 사용한 Lead-3를 baseline으로 표 3과 같이 특허 문서 5만 건에 대해 3개의 문장을 추출하고 요약문과의 Rouge Score를 측정한다. 세 가지 알고리즘은 각각 다른 방식으로 문장 간 유사도를 계산하는데 그중 Term Frequency-Inverse Document Frequency(TF-IDF) 및 코사인 유사도를 이용하는 LexRank[5]가 가장 좋은 성능을 보임을 알 수 있다.

마지막으로 생성 단계에서는 여러 태스크에서 고르게 좋은 성능을 보이며 텍스트 요약 분야에서 SOTA를 달성한 BART 기반 모델을 사용한다. 본 연구에서 사용하는 KoBART-Summarization 모델은 Text Infilling 노이즈 함수를 사용해 40GB 이상의 한국어 텍스트를 학습한 KoBART-base에 약 4만 건의 뉴스 기사 데이터 셋에 대해 사전 학습된 모델이다[6]. PatSumm 어시스턴트는 KoBART-Summarization을 활용해 앞서 가공 및 추출 단계를 거친 특허 문서 5만 건을 학습한다.

표 2. 한국어 형태소 분석기 비교 분석

특성 / 분석기	출시 년도	개발 언어	품사 태그	분석 시간	기타 특징
Kkma	2009	Java	43	1.780596	-
Mecab	2013	C/C++	43	0.007278	-
Okt	2016	Scala	19	0.206523	어간추출
Hannanum	1990	Java	29	0.270610	-
Komoran	2013	Java	45	0.060812	자소분해
Khaiii	2018	C++	46	0.072720	-
Kiwi	2018	C++	47	0.147695	-

표 3. 추출 요약 모델 비교 분석

성능 / 모델	f1 Score		
	Rouge-1	Rouge-2	Rouge-l
Lead-3	0.616840	0.523984	0.528287
TextRank	0.511778	0.368707	0.399385
LexRank	0.606330	0.507694	0.502388
Gensim	0.534207	0.396199	0.424311

4. 연구 결과

4.1 PatSumm Assistant를 활용한 웹 애플리케이션

그림 1. PatSumm 어시스턴트 개요

그림 1은 PatSumm 어시스턴트가 생성한 요약문을 출력해주는 웹 화면이다. 특허 원문을 업로드하면 PatSumm 어시스턴트는 두 가지 종류의 요약문을 제공하는데, 첫 번째는 원문 전체에 대한 내용을 담고 있는 전체 요약문이다. 전체 요약문은 사용자가 특허 문헌에 대한 주요 내용을 파악할 수 있도록 발명과 관련된 핵심 정보를 출력한다. 두 번째는 섹션 별 요약문으로 특허청에서 제공하는 요약서 작성 가이드라인에 따라 요약서에 포함되어야 하는 세 가지 항목인 기술 분야/해결과제, 해결 수단, 기대효과 각각에 대한 요약문을 출력해준다.

본 연구에서 개발한 웹 애플리케이션의 주요 기능 중 하나는 PatSumm 어시스턴트가 생성한 요약문을 편집하고 다른 곳으로 복사/붙여넣기 할 수 있는 기능이다. 요약서는 출원 시 특허 명세서와 함께 필수적으로 제출해야 하는 문서 중 하나로 특허 검색 시 활용되기 때문에 발명의 주요 내용을 담고 있어야 한다. 하지만, 작성하는 사람에 따라 양적 및 질적인 부분에서 차이가 나며 종종 요약서로 특허 문서의 핵심 내용을 파악하기 어려운 경우도 있다. PatSumm 어시스턴트는 특허청 요약서 작성 가이드를 기반으로 섹션별 요약문을 제공하기 때문에 일관된 구조의 요약서를 생성할 수 있으며, 생성된 요약문을 쉽고 빠르게 편집 및 복사/붙여넣기 할 수 있는 기능을 제공함으로써 변리사의 요약서 작성 등 업무 효율 향상에 기여한다.

4.2 성능 평가

본 연구에서는 개발한 PatSumm 어시스턴트의 요약 품질을 평가하기 위해 테스트 데이터 셋 5,000건에 대해 PatSumm 어시스턴트가 생성한 요약문과 실제 요약문 간 Rouge Score를 측정하였으며 그 결과는 표 4과 같다. 표 5는 동일한 특허에 대한 키프리스 요약문 및 PatSumm 어시스턴트를 통해 생성한 요약문을 나타낸다. 특허 문서 특성상 고정된 어휘들의 결합체가 정형화되어 관용어구처럼 사용되고 있는데 특허 명세서 및 요약서의 첫 문장에 나타나는 '본 발명은 ...에 관한 것이다.'가 그 대표적인 예이다. 표 5에서 볼 수 있듯이, 본 연구에서 제안한 PatSumm 어시스턴트가 생성한 요약문의 첫 문장도 특허 문서에서 관용적으로 사용되는 형태를 잘 반영하여 생성되었음을 확인할 수 있다. 해결 수단 부분에서도 키프리스 요약문은 제어부의 포함에 대해서만 언급하였으나 PatSumm 어시스턴트는 표시부 및 제어부의 포함을 모두 나타낸다. 또한, 키프리스 요약문에는 발명의 효과 관련 내용이 누락되어 있지만 PatSumm 어시스턴트는 특허청 요약서 작성 가이드에 구성된 세 가지 항목을 모두 포함하고 있다.

표 4. PatSumm 어시스턴트로 생성한 요약문의 Rouge Score

	Rouge-1	Rouge-2	Rouge-l
Recall	0.756119	0.644354	0.675867
Precision	0.633918	0.548914	0.568209
f1-score	0.653509	0.564305	0.586040

표 5. 키프리스 요약문-PatSumm 어시스턴트 생성 요약문 간 비교

모델	생성 요약문
키프리스 요약문	본 명세서는 터치 스크린이 잠김 상태에서도 사용자가 원하는 응용 프로그램을 용이하고 신속하게 제어할 수 있는 콘텐츠 제어 장치 및 그 방법에 **관한 것이다.** 이를 위하여 본 발명에 따른 콘텐츠 제어 장치는, 표시부의 터치 스크린이 잠김 상태일 때 응용 프로그램을 실행하기 위한 아이콘을 상기 표시부에 표시하는 **제어부**를 포함할 수 있다.
PatSumm Assistant	**본 발명은** 콘텐츠 제어 장치 및 그 방법에 **관한 것이다.** 본 발명에 따른 콘텐츠 제어 장치는, 터치 스크린을 포함하는 **표시부**와; 상기 표시부의 터치 스크린이 잠김 상태일 때 응용 프로그램을 실행하기 위한 아이콘을 상기 표시부에 표시하는 **제어부**를 포함할 수 있다. 본 발명에 의하면 터치 스크린이 잠김 상태일 때 표시부에 표시된 아이콘이 선택되면, 상기 선택된 아이콘에 링크된 응용프로그램을 실행함으로써 터치 스크린이 잠김 상태에서도 사용자가 원하는 응용 프로그램을 용이하고 신속하게 실행시킬 수 있다.

■ 기술분야/해결과제 ■ 해결수단 ■ 발명의 효과

5. 결 론

본 연구에서는 특허 문서의 핵심 내용을 파악하기 위한 특허 문서 요약 보조 시스템 PatSumm 어시스턴트를 개발했다. 긴 문서 요약 과정에서 발생하는 계산 복잡도 증가 및 노이즈 문제를 해결하기 위해 가공 및 추출 단계에서는 특허 문서의 특성을 분석하고 추출 요약 알고리즘 중 하나인 LexRank에 한국어 형태소 분석기 Mecab를 적용하였으며, 마지막 생성 단계에서는 생성 요약 모델 중 하나인 KoBART를 사용해 긴 특허 문서에 대한 요약문을 생성했다. 또한 본 연구에서 제안한 PatSumm 어시스턴트는 특허 문서에 대한 전체 요약문과 함께 특허청 요약서 작성 가이드에 나와 있는 섹션별 요약문을 함께 제공하고 편집할 수 있는 기능을 제공함으로써 특허 문서에 대한 일반인의 진입장벽 해소 및 관련 분야 종사자의 업무 효율성 제고 등을 통해 국내 특허 분야 리걸테크 시장 확산에 기여할 것으로 기대한다.

참고 문헌

[1] 김승래, "AI시대 리걸테크의 발전과 미래 법률시장의 변화 모색," *법이론실무연구*, 8(3), 2020.

[2] Gleb Sizov, "Extraction-Based Automatic Summarization: Theoretical and Empirical Investigation of Summarization Techniques." in *MS thesis, Institution for datateknikk og informasjonsvitenskap*, 2010.

[3] Mike Lewis et al., "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension." in *arXiv preprint* arXiv:1910.13461, 2019.

[4] Peter J. Liu et al., "Generating Wikipedia by Summarizing Long Sequences." in *arXiv preprint* arXiv:1801.10198, 2018

[5] Gunes Erkan and Dragomir R. Radev, "LexRank:Graph-basedLexicalCentrality as SalienceinText Summarization." in *Journal of artificial intelligence research*, 22, 2004.

[6] SK Telecom, "KoBart" [Online] Available: https://github.com/SKT-AI/KoBART (Accessed: May. 6, 2022)