

판례 내 주요정보 추출 및 추론 기술에 대한 연구

2020 年 11 月

충 남 대 학 교

공과대학 컴퓨터공학과

허 세 훈
원 준 호

목 차

I. 서론	4
II. 방법론	6
1. 데이터 셋 구축	6
1) 판례 수집 및 범죄 사실 추출	6
2) 범죄 사실에서의 주요 정보에 대한 정의	7
3) CSIE 데이터 셋	8
4) CSII 데이터 셋	9
2. 연속적 레이블링을 통한 범죄 사실 내 정보 추출	13
3. 문장 분류를 통한 범죄 사실 내 정보 추론	15
III. 관련 연구	16
IV. 실험 및 실험 결과	17
1. 연속적 레이블링을 통한 범죄 사실 내 정보 추출 실험	18
2. 문장 분류를 통한 가해자-피해자의 관계 추론 실험	20
V. 결론	23
참고문헌	23

표목차

<표 1>	7
수집한 판례에 대한 통계량	
<표 2>	8
CSIE 데이터 셋에서의 연속적 레이블별 설명, 예시, 개수	
<표 3>	11
피해자와 가해자의 관계 및 예시	
<표 4>	12
문장 분류용 데이터 셋과 문서 분류용 데이터 셋에서의 관계 태그 누적 수 및 비율	
<표 5>	14
음절 단위 BIO 태그 예시	
<표 6>	18
연속적 레이블링을 통한 범죄 사실 내 정보 추출 실험 결과	
<표 7>	20
crime.what과 crime.where에서의 의미적 청킹의 어려움	
<표 8>	22
문장 분류를 통한 가해자-피해자의 관계 분류 실험	

그림목차

<그림 1>	5
범죄 사실 내의 정보 자동 추출 및 추론 예시	
<그림 2>	12
같은 범죄 사실에 대한 문장 분류용 데이터 셋과 문서 분류용 데이터 셋의 비교	
<그림 3>	15
연속적 레이블링을 통한 문장에서 정보를 추출하는 과정	
<그림 4>	16
문장 분류를 통한 문장에서 정보를 추론하는 과정	
<그림 5>	19
한 문장당 음절의 개수에 대한 히스토그램	

I. 서 론

판례는 대표적인 전문가용 문서의 하나로서 법률적 판단을 기록해 놓은 문서이다. 법률 체계, 법률 용어의 난해함으로 인해, 전문적인 법률지식이 없는 일반인들은 판례 문서에서 원하는 정보를 얻기 힘들다. 또한 법을 잘 파악하고 있는 사람도 판례정보를 찾기 위해 많은 시간을 투자해야만 한다. 따라서 본 논문은 판례 문서를 일반인 독자가 비교적 쉽게 읽을 수 있도록 도움을 주기 위해, 판례 문서 내에서의 주요 정보를 추출하는 방법론에 대해 연구한다.

판례 문서 내에서의 사건에 대한 정보는 관련 법 조항, 법적 분쟁의 쟁점 등에 관한 정보보다 상대적으로 이해하기에 쉬운 정보이다. 판례 문서 내에서 범죄 사실은 법적 분쟁이 되는 사건들의 집합이다. 본 연구의 목적이 판례 문서에 대한 일반인 독자의 독해를 돕는 것이므로 일반인 독자가 이해하기 쉬운 정보들을 통해 정보 추출을 할 필요가 있었다. 따라서 본 논문에서는 정보 추출의 대상을 판례 문서 내에서 범죄사실에 해당하는 내용으로 한정지었다.

본 논문에서는 범죄 사실 내의 주요 정보를 자동으로 추출하기 위해서, 자연어 이해 및 자연어처리 분야에서 흔히 연구되는 연속적 레이블링(Sequence Labeling)과 문장 분류(Text Classification) 기술을 활용한다. 연속적 레이블링 문제는 주어진 입력 열(Sequence)의 각 유닛(Unit)마다 할당되어야 하는 레이블(Label)을 예측하는 문제이다. 연속적 레이블링 문제는 일반적으로 형태소 분석기나 개체명 인식기에서 많이 사용한다. 본 연구에서는 범죄 사실 내에서 추출할 수 있는 주요 정보를 여러 종류의 연속적 레이블로 정의하고 기계학습 방법을 사용하여 자동으로 정보를 추출할 수 있는 방법론을 제안한다. 또한 추출할 수는 없는 정보이지만 추론할 수 있는 정보를 정의하고 문장 분류 기술을 사용하여 자동으로 정보를 추론할 수 있는 방법론을 제안한다.

본 논문에서는 범죄 사실에 대한 데이터를 수집하고 수집된 데이터를 통해 추출 및 추론할 정보를 정의한다. 다음으로 수집한 데이터를 통해 기계학습 기반의 학습용 데이터 셋을 구축하기 위해, 정의한 정보들로 레이블링(Labeling)하는 작업을 하였고 이를 통해 연속적 레이블링 문제를 해결하기

피고인은 2011. 6. 22. 01:34경 구리시 _에 있는 ##고시텔 주방에서, 야식 준비를 하고 있던 중 평소 사이가 좋지 않던 피해자 양99(46세)가 술을 마시고 들어와 피고인에게 “씹할 너는 싸가지가 없다. 정신을 못 차린다.”라고 말을 하면서 계속하여 욕설을 하자 밖으로 나가려고 하였는데도 피해자가 몸으로 막고 못 나가게 하자 순간적으로 격분하여 피해자를 죽이기로 마음먹고, 그곳에서 햄을 자르기 위해 싱크대 위에 놓아 두었던 가위(길이 23cm, 중 제3호)를 왼손으로 잡고 피해자의 좌측 가슴 부위를 1회 힘껏 질러 피해자로 하여금 즉시 그 자리에서 대동맥손상으로 인한 다량 출혈로 사망하게 하여 피해자를 살해하였다.

- 피해자의 나이 : 46세
- 범행 일시 : 2011. 6. 22. 01:34경
- 범행 장소 : 구리시 _에 있는 ##고시텔 주방
- 범행 내용 : 살해하였다.
- 피해자와 가해자의 관계 : 이웃 주민

그림 1 범죄 사실 내의 정보 자동 추출 및 추론 예시

위해 CSIE(Case Information Extraction) 데이터 셋과 문장 분류 문제를 해결하기 위해 CSII(Case Information Inference) 데이터 셋을 구축하고 제안한다. 또한 본 논문에서는 제안한 데이터 셋들을 활용하여 범죄 사실 내의 정보 자동 추출 및 추론을 위한 방법론을 제안한다. 제안하는 방법론을 통해 추출 및 추론을 진행한 결과는 그림 1과 같다.

최근 자연어 이해 및 처리 분야는 심층신경망 기반의 학습 방법(Deep-Learning)들을 통해 혁혁한 성과를 보여주었고[1, 2, 3] 현재 대부분의 심층 신경망(Deep Neural Network) 접근법들은 BERT(Bidirectional Encoder Representations from Transformers)[1]를 기반으로 한 모델들[4, 5]이 활발히 연구되고 있다. 한국어의 경우에도 SKTBrain에서 배포한 KoBERT[6]를 비롯해, ELECTRA(Efficiently Learning an Encoder that Classifies Token Replacements Accurately)[4] 모델을 변형해 한국어의 특성을 반영한 KoELECTRA[7] 등이 연구되었다. 본 논문에서는 Multilingual BERT, KoBERT, KoELECTRA를 통한 범죄 사실 내의 정보 추출 및 추론 실험을 진행한다.

본 논문의 2장은 제안하는 데이터 셋을 구축하는 과정을 기술하고 구축한 데이터 셋을 활용하여 정보 추출 및 추론을 자동으로 할 수 있는 방법론을

제안한다. 3장에서는 본 논문과 관련된 연구들을 살펴보고 본 연구와의 차이점을 기술한다. 4장에서는 실험을 통해 3장에서 제안한 방법론에 대해 검증하고 결과를 분석한다. 5장에서는 본 연구의 결론과 제안한 방법론에 대한 활용 가능성 및 향후 연구에 대해 언급한다.

II. 방법론

본 장에서는 정성적 분석을 통해 판례 문서 내에서 추출할 정보와 추론할 정보를 정의한 후에 CSIE 데이터 셋과 CSII 데이터 셋을 구축한다. 다음으로 연속적 레이블링 문제를 적용하여 정보를 추출하는 방법론과 문장 분류 기술을 적용하여 정보를 추론하는 방법론을 제안한다. 본 장의 마지막에서는 판례 문서 내에서 문장 분류를 통해 정보를 추론하는 것의 문제점을 언급하고 이를 개선하기 위한 방법론을 제안한다.

1. 데이터 셋 구축

본 논문에서는 한국어 판례를 수집하고 각 판례에서 범죄 사실에 해당하는 내용을 추출하였다. 판례 내의 범죄 사실은 피고인이 저지른 범행에 대한 구체적인 사실을 기술한 부분이며, 법적 분쟁이 되는 사건들의 집합이다. 본 연구는 판례 문서에 대한 일반인 독자의 독해를 돕는 것을 목적으로 하므로 일반인 독자가 이해하기 쉬운 정보들을 통해 정보 추출을 할 필요가 있었다. 범죄 사실은 판례 내의 다른 부분들에 비해, 상대적으로 일반인 독자가 이해하기에 쉽다. 따라서 본 논문에서는 데이터 셋 구축을 위한 수집 대상을 판례 문서 내에서 범죄 사실로 한정하였다. 본 논문에서는 범죄 사실에 대한 주요정보를 정의하고 이에 근거하여 CSIE 데이터 셋과 CSII 데이터 셋을 구축하였다.

1) 판례 수집 및 범죄 사실 추출

본 논문에서는 법률정보통합검색 사이트인 ‘리걸서치’[8]에서 성범죄 관련 판례 251건, 폭행 관련 판례 155건, 살인 관련 판례 30건을 수집하였다. 최종적으로는 수집한 판례에서 두 가지 이상의 유형이 함께 나와서 중복된 경우

수집한 판례 문서 수(개)	402
판례 내의 평균 범죄 사실 수(개)	2.26
수집한 총 문장 수(개)	1595
각 문장의 평균 글자 수(개)	173.64
수집한 문장에서의 최대 글자 수(개)	955

표 1 수집한 판례에 대한 통계량

를 제거하여 총 402건을 수집하였다.

판례 문서는 1심, 2심, 3심에 따라, 그 구조가 다르다. 일반적으로 2심과 3심 판례 문서는 1심 판례 문서를 기반으로 작성되기 때문에 범죄 사실이 기술되어 있지 않다. 따라서 본 논문에서는 수집 대상이 되는 판례를 1심 재판에 대한 판례로만 한정하였다. 또한 본 연구에서는 범죄 사실에 대한 주요정보를 자동으로 추출하는 것이 목표이므로 범죄 사실이 명시되어 있지 않은 판례는 수집에서 제외하였다.

수집한 판례는 심층신경망의 입력 토큰 수가 제한되어 있는 경우 등을 고려해서 문장 구분을 해주었다. 이 때, 문장 구분은 3인의 작업자에 의해 행해졌다. 표 1은 수집한 판례 문서 수, 판례 내의 평균 범죄 사실 수, 총 문장 수, 각 문장의 평균 글자 수를 보여준다.

2) 범죄 사실에서의 주요 정보에 대한 정의

판례 내의 범죄 사실은 사건에 대한 경위가 서술되어 있는 부분이다. 사건은 흔히 육하원칙에 맞게 서술할 수 있다. 따라서 범죄 사실도 “누가?”, “언제?”, “어디서?”, “무엇을?”, “어떻게?”, “왜?”의 여섯 가지 구조로 구성된다고 가정할 수 있다. 하지만 “어떻게?”와 “왜?”에 관한 정보들은 그 범위가 너무 다양하고 정보를 추출할 수 없는 문장들이 대부분이었다. “누가?”에 관한 정보들은 “피고인 1”, “위 피고인”과 같이 정보를 추출할 수는 있지만 추출된 정보에서 의미적인 정보량이 없는 경우가 대부분이었다. 또한 본 논문에서 수집한 판례는 성범죄 관련 판례가 60퍼센트 이상을 차지한다. 성범죄의 경우, 피해자의 나이에 따라, 범죄의 경중이 달라지는 경우가 많다. 따라서 본 연구에서는 “범행 시기”, “범행 장소”, “범행 내용”, “피해자의 나이”를 범죄

사실에 대한 주요 정보로 정의하였다.

한편, 본 논문에서는 사건에 대한 경위뿐만 아니라, 피해자와 가해자간의 관계에 대한 정보도 범죄 사실에서는 주요 정보가 될 수 있다고 판단하였다. 예를 들어, 성폭력범죄의 처벌 등에 관한 특례법 제 5조(친족관계에 의한 강간 등)에 관한 법률에 따르면, 친족 관계라는 이유로 가중처벌이 가능하다고 명시되어 있다. 이에 본 논문에서는 범죄 사실에서 추론할 수 있는 피해자-가해자의 관계도 주요 정보로 정의하였다.

3) CSIE 데이터 셋

CSIE 데이터 셋은 앞선 장에서 정의한 범죄 사실에 대한 주요 정보 중에 “범행 시기”, “범행 장소”, “범행 내용”, “피해자의 나이”의 4가지 정보들을 위한 데이터 셋이다. [9]에서는 이러한 정보를 기계독해 기술을 활용하여 자동 추출하였다. 기계독해 기술은 주어진 지문과 질문을 이해하여 지문 내에서 답변을 찾는 기술이다. [9]에서 제안한 방법론을 활용하기 위해서는 질문에 대한 질문-답변 쌍이 항상 존재해야만 한다. 하지만 앞서 추출하기로 결정한 정보들은 존재할 수도 있고 존재하지 않을 수도 있다. 따라서 이러한

연속적 레이블 이름	설명	예시	훈련 데이터 셋에 대한 개수	평가 데이터 셋에 대한 개수	전체 데이터 셋에 대한 개수
crime.when	범행 일시	2013. 6. 초순 21:00경	732	197	929
crime.where	범행 장소	피해자 고△△의 거주지 현관	696	191	887
crime.what	범행 내용	추행하였다.	1808	513	2321
victim.age	피해자의 나이	30세	440	109	549

표 2 CSIE 데이터 셋에서의 연속적 레이블별 설명, 예시, 개수

정보들이 항상 모든 문장에 나와야만 한다고 가정하고 데이터 셋을 구축하는 것은 옳바르지 않다. 또한 [9]는 질문-답변 쌍이 1:1로 대응되는 구조이므로 한 지문 내에서의 답변이 여러 개일 경우, 추출할 수 없다는 점이 있다. 따라서 본 논문에서는 “범행 시기”, “범행 장소”, “범행 내용”, “피해자의 나이” 정보를 추출하기 위해 각 정보를 하나의 연속적 레이블로 할당한다. 즉, CSIE 데이터 셋은 연속적 레이블링을 위한 학습 데이터와 평가 데이터로 구성되는 데이터 셋이다. 표 2는 CSIE 데이터 셋의 연속적 레이블의 이름, 설명, 예시와 데이터 셋에서의 개수를 보여준다.

4) CSII 데이터 셋

CSII 데이터 셋은 피해자-가해자의 관계에 대한 정보를 사전에 정의한 관계로 추론하기 위한 데이터 셋이다. 본 논문에서는 피해자-가해자의 관계 정보를 추론하는 문제를 단순 지도 학습(Supervised Learning) 기반의 문장 또는 문서 분류 문제로 변환한다. 따라서 본 논문에서는 CSII 데이터 셋을 구축하기 위해 피해자와 가해자의 관계를 표 3과 같이 사전에 정의하였다.

하지만 범죄 사실 내에서 피해자-가해자 관계 정보는 한 문장으로 판단하기 어려운 경우가 많다. 또한 문장 분류를 통해 가해자-피해자의 관계 정보를 정확히 분류해내어도 그 결과는 단순히 특정 문장에 대한 분류 결과이다. 즉, 판례 문서 전체에서의 피해자-가해자의 관계가 분류되는 것은 아니다. 따라서 본 논문에서 제안하는 CSII 데이터 셋은 문장 분류용 데이터 셋과 문서 분류용 데이터 셋으로 구성되며, 각각의 분류용 데이터 셋은 학습 데이터와 평가 데이터로 구성된다. 또한 모델이 문장에 대해 예측한 결과를 문서에 대한 예측 결과와 비교할 수 있도록 각각의 분류용 데이터 셋은 동일한 판례 문서들로 구성하였다. 즉, 문장 분류용 데이터 셋은 문서 분류용 데이터 셋에서 문장별로 분리하여 문장 단위 내에서의 가해자-피해자의 관계 태그를 부착한 것이다. 그림 2는 동일한 판례 문서에 대한 문장 분류용 데이터 셋과 문서 분류용 데이터 셋의 실제 예를 보여준다. 그림에서 확인할 수 있듯이 문장 분류용 데이터 셋은 문서 분류용 데이터 셋에서 문장별로 분리하여 가해자-피해자의 관계 태그를 부착한 것이다. 표 4는 피해자-가해자의 관계에 대한 문장 분류용 데이터 셋(D_1)에서의 누적 수, 비율과 문서 분류용 데이터

셋(D_2)에서의 누적 수와 비율을 보여준다.

관계	예시 문장
부부	피고인은 유통업에 종사하는 사람으로서, 피해자 조□□(여, 37세)의 남편이다.
연인	피고인은 2011. 6. 24. 대구 서구 중리동 달서초등학교 부근 노상에서 피해자가 헤어지자고 했다는 이유로 위 택시를 운행하면서 운전석 쪽 뒷바퀴로 그녀의 오른발을 충격하여 피해자에게 치료기일 불상의 발 부분의 염좌 및 긴장(우측) 등의 상해를 가하였다.
친족관계	피고인은 평소 함께 살고 있는 어머니인 피해자 강○○가 술을 많이 마신다는 이유로 강한 불만을 갖고 있었다.
동료/동업	피고인과 피해자 ○○○(여, 19세)는 충주시 성서동에 있는 △△△미술학원의 강사이다.
손님-점원	피고인은 택시면허가 없이 D 영업용 택시를 운전하던 중 2016. 8. 9. 08:29경 대구 중구에 있는 봉산육거리 부근에서 피해자 E(가명, 여, 19세)을 F으로 태우게 되었다.
스승-제자	피고인은 중학교 체육교사로 2007. 3.경부터 2010. 3.경까지 ○○중학교의 육상부 감독으로 근무하면서 육상부원이었던 청소년 피해자 김○○(15세)를 지도하였고, 다른 학교로 전근한 이후에도 피해자를 만나 상담한 적이 있었다.
이웃주민	피고인은 서울 강서구 화곡로 18길에 있는 건물 **호에 사는 사람이고, 피해자 B은 같은 건물 00호에 사는 이웃 주민이다.
지인	피고인은 2014. 4. 7. 19:50경 수원시 인계동에 있는 ‘□□’이라는 술집에서 피해자 최○○(여, 28세)을 비롯한 ‘맛있는 ☆☆’이라는 인터넷 카페 회원들과 술을 마시다가 ‘♣♣ 노래방’으로 자리를 옮겨 계속하여 술을 마셨다.

군대 관련	피고인은 2016. 1. 하순 22:00경 춘천시 C에 있는 국군00병원 정신과병동 간호장교실에서, 같은 소속 의무병으로 근무하는 일병인 피해자 D(20세)에게 환자용 억제대 착용법을 알려준다며 의자에 앉히고 피해자의 양 손목과 발목에 위 억제대를 채워 움직이지 못하게 한 후, 피해자의 양 무릎을 잡아 강제로 피해자의 다리를 벌리고 손가락으로 피해자의 성기를 4회 튕겨 때렸으며, 손으로 피해자의 왼쪽 가슴을 움켜잡아 비틀고 피해자의 성기를 2회 잡아 비틀었다.
종교 관련	피고인은 목포시에 거주하고 있는 불교신자인 피해자 도○○(여, 28세)이 전화로 “남편이 화상을 입는 등 우환이 있다.”고 하자, 현재 자신이 머무르고 있는 경남 산청군 시천면에 있는 지리산으로 오라고 하였다.
형사-피의자	피고인은 2014. 5. 21. 03:20경 울산남부경찰서 옥동지구대에서 2014. 5. 15.경 발생했던 피고인의 폭행 사건 당시 작성된 피고인에 대한 현행범인체포서에 기재된 범죄사실이 피고인의 기억하는 내용과 다르다며 항의를 하던 중 흥분을 하여 위 지구대에서 근무하는 경찰관들에게 "썩할놈들 건드리지마라, 건드리면 죽여버리고 감옥가려고 왔다"라고 욕설을 하고 소란을 피우던 중, 순찰근무를 마치고 지구대 안으로 들어온 경위 C과 경장 김광섭이 피고인에게 욕설을 하지 말라고 했다는 이유로 "이썩할놈 나를 잡아넣으려고 몰려오네"라고 외치며 피고인의 머리로 위 C의 가슴 부분을 1회 들이받고, 오른손 주먹으로 C의 왼쪽 목 부분을 1회 때려 폭행하였다.
낮선 사람	피고인은 2013. 7. 4. 23:35경 경산시 중방동에 있는 경산시장 내 금은방 앞 노상에서 그랜저 XG 승용차를 운행하다가, 파지를 수집하고 있던 피해자(여, 69세)를 발견하고 강간하기로 마음먹은 후, 다시 위 차량을 운전하여 위 피해자가 있는 위 경산시장 내 ‘마이하우스’ 분식점 앞으로 이동하였다.
알 수 없음	그 후 피고인은 그곳에 놓여있던 피해자 소유의 10만 원 권 자기 앞수표 1장, 현금 24만원, 신용카드 4장 등이 들어 있던 손가방을 들고 나갔다.

표 3 피해자와 가해자의 관계 및 예시

<p>피고인 겸 피부착명령청구자(이하 '피고인'이라 한다)는 부산 동래구 ●●동 소재 ●●●● 아파트 ●●동 ●●호에 거주하고, 피해자 이●●(여, 10세)는 피고인의 바로 옆집에 거주하고 있다. - 이웃주민</p>	<p>피고인 겸 피부착명령청구자(이하 '피고인'이라 한다)는 부산 동래구 ●●동 소재 ●●●● 아파트 ●●동 ●●호에 거주하고, 피해자 이●●(여, 10세)는 피고인의 바로 옆집에 거주하고 있다. 피고인은 피해자의 부모가 맞벌이를 하기 때문에 낮 시간에는 피해자와 피해자의 남동생 이●●(남, 8세)만 집에 있다는 것을 알고 피해자를 피고인의 집으로 유인하여 추행할 것을 마음먹었다. 피고인은 2012. 7. 27. 13:00경 피고인의 집 앞 복도에서 피해자와 이●●이 피해자의 집 현관문을 열고 집으로 들어가는 것을 발견하고, 이●●에게는 집으로 들어가라고 말하고, 피해자에게 "집에 아이스크림이 있으니 주겠다"고 말하고 피해자를 피고인의 집으로 데리고 왔다. - 이웃주민</p>
<p>피고인은 피해자의 부모가 맞벌이를 하기 때문에 낮 시간에는 피해자와 피해자의 남동생 이●●(남, 8세)만 집에 있다는 것을 알고 피해자를 피고인의 집으로 유인하여 추행할 것을 마음먹었다. - 이웃주민</p>	<p>피고인은 2012. 7. 27. 13:00경 피고인의 집 앞 복도에서 피해자와 이●●이 피해자의 집 현관문을 열고 집으로 들어가는 것을 발견하고, 이●●에게는 집으로 들어가라고 말하고, 피해자에게 "집에 아이스크림이 있으니 주겠다"고 말하고 피해자를 피고인의 집으로 데리고 왔다. - 이웃주민</p>
<p>피고인은 냉창고에서 아이스크림 3개, 소시지 1개를 꺼낸 후 피해자에게 "맛을 보고 무엇인지 알아 맞추기 게임을 하자"고 말한 후 수건으로 피해자의 눈을 가려놓아 피해자를 방심하게 한 다음, 자신의 성기를 꺼내어 피해자의 입에 집어넣었다. - 알 수 없음</p>	<p>피고인은 냉창고에서 아이스크림 3개, 소시지 1개를 꺼낸 후 피해자에게 "맛을 보고 무엇인지 알아 맞추기 게임을 하자"고 말한 후 수건으로 피해자의 눈을 가려놓아 피해자를 방심하게 한 다음, 자신의 성기를 꺼내어 피해자의 입에 집어넣었다. - 이웃주민</p>

그림 2 같은 범죄 사실에 대한 문장 분류용 데이터 셋과 문서 분류용 데이터 셋의 비교

	D_1 에서의 누적 수(개)	D_1 에서의 비율(%)	D_2 에서의 누적 수(개)	D_2 에서의 비율(%)
부부	29	0.018	12	0.029
연인	12	0.008	5	0.012
친족관계	159	0.100	37	0.092
동료/동업	74	0.046	29	0.072
손님-점원	64	0.040	42	0.104
스승-제자	67	0.042	13	0.032
이웃주민	18	0.011	9	0.022
지인	235	0.147	88	0.218
군대 관련	23	0.014	6	0.014
종교 관련	8	0.005	3	0.007
형사-피의자	139	0.087	7	0.017
낯선 사람	296	0.186	136	0.338
알 수 없음	471	0.295	15	0.037
합계	1595	1.000	402	1.000

표 4 문장 분류용 데이터 셋(D_1)과 문서 분류용 데이터 셋(D_2)에서의
관계 태그 누적 수 및 비율

2. 연속적 레이블링을 통한 범죄 사실 내 정보 추출

2.1장에서 범죄 사실 내에서 정보를 추출할 수 있는 경우와 추론할 수 있는 경우를 정의하였다. 본 장에서는 정보를 추출해야만 하는 경우에 대한 방법론을 제안한다.

연속적 레이블링 문제는 주어진 입력 열(Sequence)의 각 유닛(Unit)마다 할당되어야 하는 레이블(Label)을 예측하는 문제이다. 연속적 레이블링 문제 내의 대표적인 분야에는 개체명 인식(Named Entity Recognition) 기술이 있다. 개체명 인식은 정보 추출의 한 분야로서 문서 내에서 개체명(Named Entity)을 추출하고 추출된 개체명의 종류를 결정하는 작업을 말한다. 기계학습 기반의 개체명 인식 기술들은 주로 BIO 태그[10]를 이용하는데, 본 논문에서는 음절 단위의 각 문자에 BIO 태그를 부착한다. 또한 레이블링된 각 문자는 심층신경망의 입력 열의 각 유닛으로 사용하고 BIO 태그는 심층신경망의 레이블로 사용하였다. 표 5는 “피고인은 2013. 10. 8. 11:30 경 위 문구점에서, 피해자 김■■■(여, 14세)에게 ...” 문장을 BIO 태그로 표현한 예이다.

음절	BIO 태그	음절	BIO 태그
피	O	위	O
고	O		O
인	O	문	B-crime.where
은	O	구	I-crime.where
	O	점	I-crime.where
2	B-crime.when	에	O
0	I-crime.when	서	O
1	I-crime.when	,	O
3	I-crime.when		O
.	I-crime.when	피	O
	I-crime.when	해	O
1	I-crime.when	자	O
0	I-crime.when		O
.	I-crime.when	김	O
	I-crime.when	■	O
8	I-crime.when	■	O
.	I-crime.when	(O
	I-crime.when	여	O
1	I-crime.when	,	O
1	I-crime.when		O
:	I-crime.when	1	B-victim.age
3	I-crime.when	4	I-victim.age
0	I-crime.when	세	I-victim.age
경	I-crime.when)	O
	O

표 5 음절 단위 BIO 태그 예시

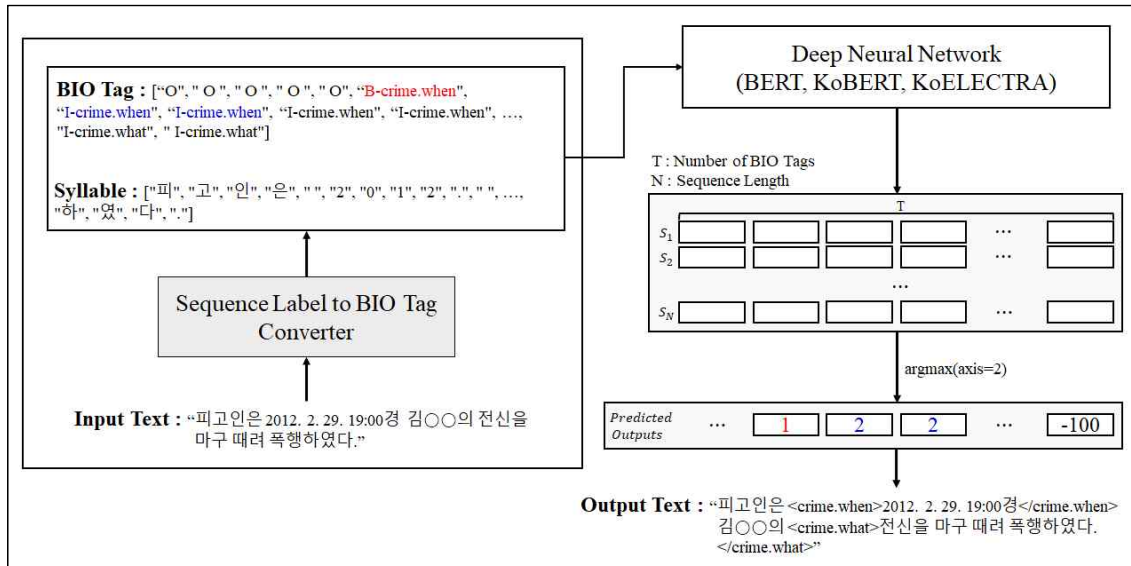


그림 3은 연속적 레이블링을 수행하여 문장에서 정보를 추출하는 과정을 보여준다. 그림 3은 “피고인은 2012.2.29. 19:00경 김○○의 전신을 마구 때려 폭행하였다.”라는 문장이 BIO 태그 변환기를 통해 음절과 BIO 태그로 변환되고 이들이 심층신경망의 입력으로 사용됨을 보여준다. 심층신경망은 입력 열의 개수(N)에 대해 BIO태그의 수(T)만큼의 벡터를 결과로 내놓는다. 심층신경망의 출력은 가장 높은 확률을 가지는 BIO 태그의 인덱스 값으로 변환된다. 변환된 인덱스 값은 입력 문장에 대해서 정보 추출 태그가 부착된 형태로 변환된다. 본 연구에서는 이러한 과정으로 심층신경망을 학습시키고 입력 열에 대한 연속적 레이블의 예측을 수행하여, 범죄 사실 내에서 정보 추출을 하였다.

3. 문장 분류를 통한 범죄 사실 내 정보 추론

본 장에서는 범죄 사실 내에서 정보를 추론할 수 있는 경우에 대한 방법론을 제안한다.

2.1장에서는 범죄 사실 내에서 피해자-가해자 간의 관계를 정의하였다. 따라서 본 논문에서는 판례 전체에서 피해자와 가해자의 관계가 사전에 정의

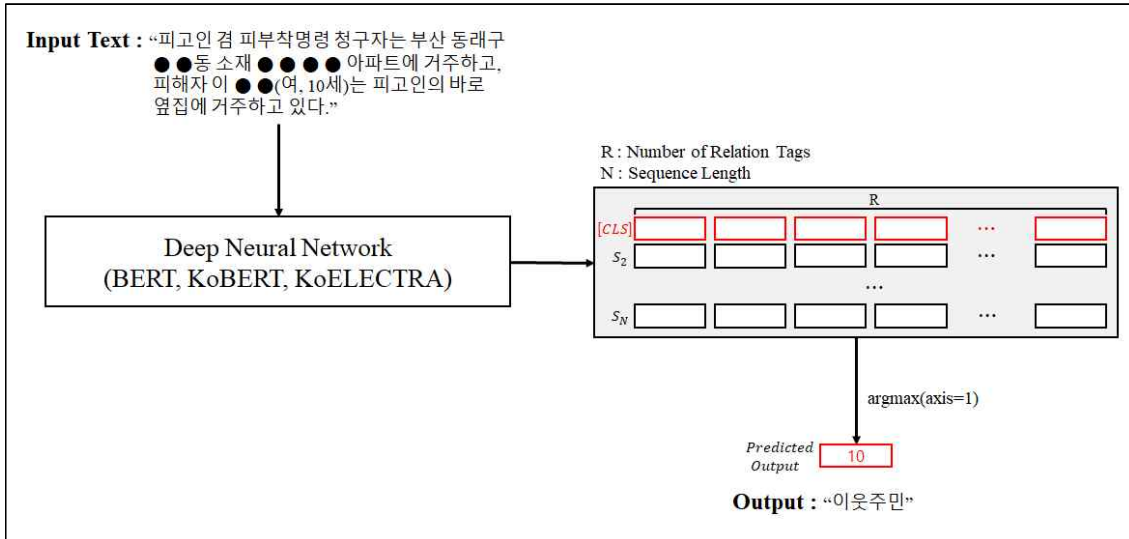


그림 4 문장 분류를 통한 문장에서 정보를 추론하는 과정

한 관계 중 하나로만 추론될 수 있음을 가정한다. 이러한 가정을 통해 본 장에서는 자연어 처리 분야에서 보편적으로 연구된 문장 분류 기술을 사용한다. 그림 4는 “피고인 겸 피부착명령 청구자는 부산 동래구 ●●동 소재 ●●●아파트에 거주하고, 피해자 이●●(여, 10세)는 피고인의 바로 옆집에 거주하고 있다.”라는 문장이 심층신경망에 입력되어 문장 분류를 위한 의미 벡터인 [CLS] 벡터를 통해 관계 태그의 인덱스 값으로 변환되는 과정을 보여준다.

Ⅲ. 관련 연구

전문가 수준의 문서를 분석하고 이를 통해 정보를 얻는 것은 일반인들에게 대단히 어려운 일이다. 예를 들어, 특허 지식이 없는 일반인이 선행기술을 조사하고 본인의 특허와 유사한지를 분석하는 것은 어려우며, 법률지식이 없는 일반인이 변호사의 도움 없이 본인이 처한 어려움에 해당하는 법령을 찾고 활용하는 것 역시 매우 힘든 일이다.

전문가 수준의 문서를 분석하려는 연구는 활발하게 진행되었다. [11]은 특허 문서 내의 단어들에 대한 TF-IDF 점수와 IPC 코드를 통해 특허 문서를

분석하고 각 문서에 대한 간단한 포토폴리오를 제공하였다. [12]에서는 생의학 분야의 비정형 텍스트에서 핵심적 내용을 추출할 수 있는 기계학습 기반 정보 추출 시스템을 구축하기 위한 언어자원 수집 및 통합적 구조화 방안을 제안하였다.

판례 문서에서도 정보를 추출하여 분석하려는 연구가 있었다. [13]은 TextRank[14]를 활용하여 문서를 요약하고 이를 선학습된 BERT 모델[1]의 입력으로 사용하여 문서 요약 판례 문서 검색 시스템의 성능 향상을 도모하였다. [15]에서는 판례 문서에서의 개체명 체계를 정의하고 개체명 인식 기술을 활용하여 판례 문서 내 핵심 정보를 추출하였다. [9]에서는 기계 독해를 이용하여 판례 내 주요 정보를 추출할 수 있는 데이터 셋과 모델을 제안하였다.

이전 연구의 방법들과는 달리 본 논문에서는 주요 정보 추출뿐만 아니라, 판례 문서 내에서 얻을 수 없는 정보도 정의하고 추론한다. 정보 추론을 위한 대부분의 연구들은 목적에 맞는 온톨로지(Ontology)를 구축하고 온톨로지 규칙 기반의 추론 알고리즘을 구현하려는 방식으로 행해졌다[16]. 하지만 온톨로지를 구축하는 작업은 많은 시간과 비용을 필요로 한다. 따라서 본 논문은 온톨로지를 구축하지 않고 모델 기반의 자동 문장 및 문서 분류 방법론을 취하여 정보를 추론할 수 있음을 증명한다.

문장 분류 방법론은 자연어 처리 분야에서 활발하게 연구되었다. [1]에서는 선학습시킨 모델에 전이 학습을 적용하여 다양한 테스트들로 구성된 GLUE 벤치마크[17]에서 SOTA 성능을 기록하였다. 본 논문에서는 BERT 기반의 문장 분류 모델을 사용하여 정보 추론을 하는 모델을 제안한다.

IV. 실험 및 실험 결과

본 연구의 목적은 판례 문서의 범죄 사실 내에서 주요 정보를 추출하는 것이다. 또한 2.1.4장에서 밝혔듯이 추론할 수 있는 정보도 있다. 본 장에서는 제안한 데이터 셋들은 활용하여 자동으로 정보 추출과 추론을 하는 실험들을 수행한다. 본 장의 마지막에서는 실험에서의 오류를 지적하고 이를 개선할 수 있는 기법을 제안하고 실험을 수행한다.

1. 연속적 레이블링을 통한 범죄 사실 내 정보 추출 실험

본 장에서는 범죄 사실 내에서 주요정보 추출이 올바르게 수행되는지 검증하기 위해 연속적 레이블링을 통한 정보 추출 실험을 진행한다. 실험에 사용한 데이터 셋은 2.1.3장에서 제안한 CSIE 데이터 셋이고 BIO 태그로 레이블링된 문장을 신경망의 입력으로 사용하고 crime.when, crime.where, crime.what, victim.age의 총 4가지의 연속적 레이블을 예측하도록 하였다. 실험에 사용한 모델은 Multilingual BERT, KoBERT, KoELECTRA이고 실험 결과는 표 6과 같다.

판례 문서는 기본적으로 문장을 길게 늘어서 쓰는 경우가 많다. 실제로 표 1에서 확인할 수 있듯이 본 논문에서 수집한 데이터 셋에서 문장 내 음절의 최대 수는 955개이다. 2.2장에서 연속적 레이블링을 위해 본 연구에서는 심층신경망의 각 입력 단위로 음절 하나를 사용한다고 하였다. 하지만 실험에

	연속적 레이블	Precision	Recall	F1
Multilingual BERT	crime.what	0.447	0.631	0.524
	crime.when	0.758	0.862	0.807
	crime.where	0.191	0.219	0.204
	victim.age	0.862	0.917	0.888
	Total	0.565	0.658	0.606
KoBERT	crime.what	0.080	0.207	0.116
	crime.when	0.740	0.852	0.792
	crime.where	0.248	0.319	0.279
	victim.age	0.636	0.834	0.722
	Total	0.426	0.553	0.477
KoELECTRA	crime.what	0.415	0.653	0.507
	crime.when	0.728	0.873	0.794
	crime.where	0.139	0.193	0.162
	victim.age	0.877	0.917	0.896
	Total	0.540	0.659	0.590

표 6 연속적 레이블링을 통한 범죄 사실 내 정보 추출 실험 결과

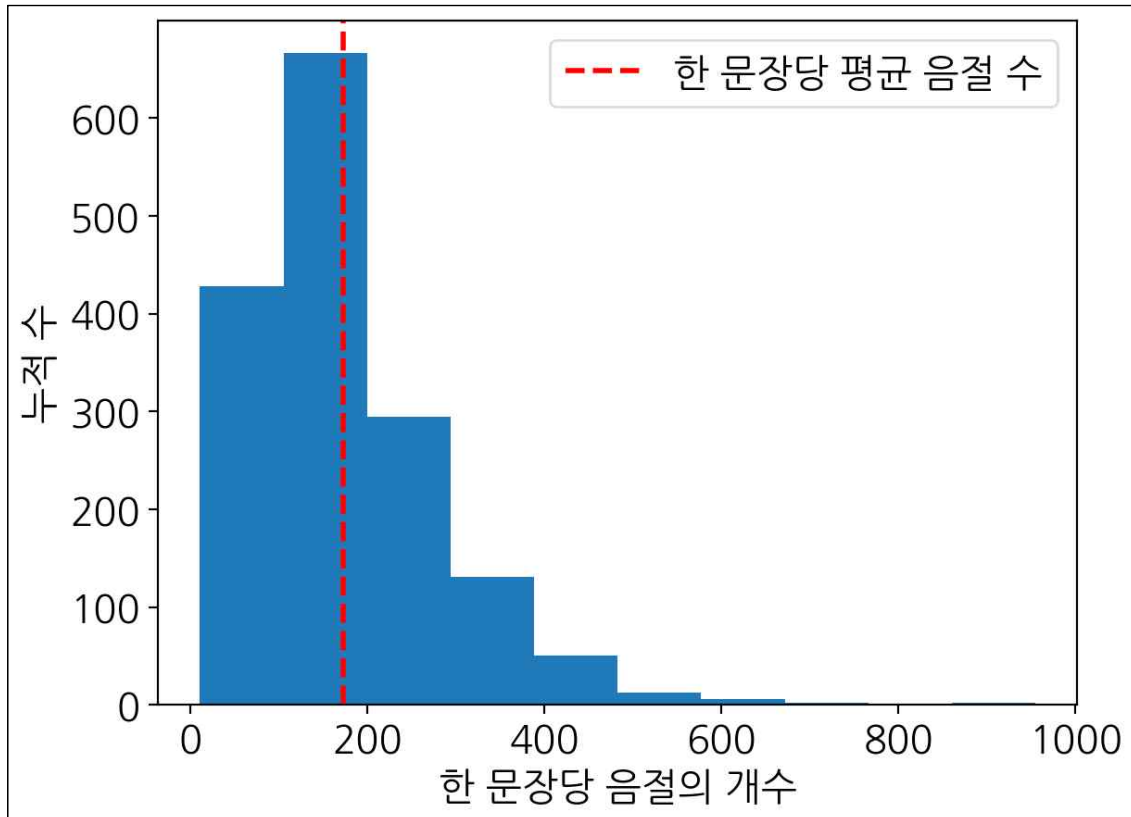


그림 5 한 문장당 음절의 개수에 대한 히스토그램

서 사용하는 BERT 기반의 모델들은 최대 토큰 수가 512개로 제한된다. 따라서 본 연구에서는 불가피하게 문장 내 최대 음절수가 512개를 넘어설 경우, 나머지 입력 열은 사용하지 않도록 실험을 설계했다. 또한 그림 5를 보면 대부분의 문장은 평균 음절수에 편향되어 있기 때문에 전체적인 성능에는 영향을 미치지 않을 것이라고 가정하고 실험을 진행하였다.

표 6을 보면 `crime.what`과 `crime.where`은 상대적으로 실험 성능이 좋지 않음을 확인할 수 있다. `crime.what`과 `crime.where`의 경우, 의미의 단위로 올바르게 청킹(Chunking)하기가 어렵다. 반면, `crime.when`과 `victim.age`의 경우, 범죄 사실 내에서 출현하는 의미 단위가 거의 정해져 있다. 표 7은 `crime.what`과 `crime.where` 연속적 레이블의 실제 예시 문장을 통해 의미 단위로 청킹하기 어려움을 보여준다. 표 7에서는 동일한 문장에 대해, 각 연속적 레이블에 대한 정보를 세 가지 기준에 따라 추출한 예시를 보여준다. 이를 통해 미루어

연속적 레이블	예시 문장	기준 번호	추출된 정보
crime.what	위험한 물건인 칼로 피해자 신oo를 찌를 듯이 위협하면서 다른 손으로 피해자 신oo의 얼굴을 3회 때렸다.	기준 1	다른 손으로 피해자 신oo의 얼굴을 3회 때렸다.
		기준 2	피해자 신oo의 얼굴을 3회 때렸다.
		기준 3	얼굴을 3회 때렸다.
crime.where	피고인은 울산 울주군 건물 2층에 있는 피해자 B 운영의 ‘C다방’에서 피해자의 머리 부위를 수회 내리쳤다.	기준 1	울산 울주군 건물 2층에 있는 피해자 B 운영의 ‘C다방’
		기준 2	피해자 B 운영의 ‘C다방’
		기준 3	C다방

표 7 crime.what과 crime.where에서의 의미적 청킹의 어려움

볼 수 있듯이, crime.what과 crime.where에서는 적절한 의미 단위로 청킹하는 것이 어렵기 때문에 상대적으로 실험 성능이 낮은 이유를 추측할 수 있다.

2. 문장 분류를 통한 가해자-피해자의 관계 추론 실험

본 장에서는 범죄 사실 내에서 가해자-피해자의 관계 정보 추론이 올바르게 수행되는지 검증하기 위해 문장 분류 실험을 진행한다. 실험에 사용한 데이터 셋은 2.1.4장에서 제안한 CSII 데이터 셋이고 한 문장을 신경망의 입력으로 사용하고 2.1.4장에서 정의한 관계 중 하나로 문장을 분류하도록 하였다. 실험에 사용한 모델은 4.1장과 같다. 표 8은 사용한 모델과 가해자-피해자의 관계에 따른 정확도와 F1 성능과 전체 관계 클래스(Class)에 대한 성능을 보여준다.

	관계 클래스	정확도	F1
Multilingual BERT	부부	0.250	0.400
	연인	0.000	0.000
	친족관계	0.987	0.857
	동료/동업	0.636	0.250
	손님-점원	0.176	0.240
	스승-제자	0.315	0.480
	이웃주민	0.000	0.000
	지인	0.468	0.389
	군대 관련	0.000	0.000
	종교 관련	0.000	0.000
	형사-피의자	0.915	0.864
	낮선 사람	0.655	0.666
	알 수 없음	0.630	0.681
	Total	0.599	0.597

KoBERT	부부	0.000	0.000
	연인	0.000	0.000
	친족관계	0.687	0.687
	동료/동업	0.363	0.275
	손님-접원	0.294	0.454
	스승-제자	0.421	0.551
	이웃주민	0.000	0.000
	지인	0.404	0.380
	군대 관련	0.000	0.000
	종교 관련	0.000	0.000
	형사-피의자	0.966	0.966
	낯선 사람	0.819	0.675
	알 수 없음	0.739	0.727
	Total	0.650	0.631
KoELECTRA	부부	0.000	0.000
	연인	0.000	0.000
	친족관계	0.937	0.638
	동료/동업	0.181	0.148
	손님-접원	0.058	0.111
	스승-제자	0.315	0.428
	이웃주민	0.000	0.000
	지인	0.489	0.410
	군대 관련	0.000	0.000
	종교 관련	0.000	0.000
	형사-피의자	0.949	0.918
	낯선 사람	0.819	0.671
	알 수 없음	0.621	0.678
	Total	0.610	0.583

표 8 문장 분류를 통한 가해자-피해자의 관계 분류 실험

V. 결론

본 논문에서는 판례 문서를 일반인 독자가 비교적 쉽게 읽을 수 있도록 도움을 주기 위해, 판례 문서 내에서의 주요 정보를 정의하고 주요 정보 추출을 위한 CSIE 데이터 셋과 주요 정보 추론을 위한 CSII 데이터 셋을 제안하였다. 또한 본 논문에서는 제안한 데이터 셋을 활용하여 주요 정보를 추출하는 방법론과 추론하는 방법론을 제안하고 실험을 통해 방법론을 검증하였다.

본 논문에서 제안한 추출 및 추론 방법을 활용하면 범죄 사실 내에서의 사건 요약 문장 생성, 판례 문서 내의 주요 정보 하이라이트 기능, 일반인 사용자 맞춤 판례 검색 시스템을 개발할 수 있을 것이다. 향후에는 일반인뿐만 아니라, 법조인들을 위한 주요 정보들을 정의하고 판례 문서의 전체적인 구조를 포괄하는 체계를 제안할 수 있는 연구를 진행할 것이다.

참고문헌

- [1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171 - 4186, 2019.
- [2] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut, "Albert: A lite bert for self-supervised learning of language representations," In Proceedings of the 2020 Conference of the International Conference on Learning Representations, 2020.
- [3] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," arXiv preprint arXiv:1910.10683, 2019.
- [4] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators", arXiv preprint arXiv:2003.10590, 2020.

v:2003.10555, 2020.

- [5] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter”, arXiv preprint arXiv:1910.01108, 2020.
- [6] SKTBrain, “Korean BERT pre-trained cased (KoBERT),” <https://github.com/SKTBrain/KoBERT>
- [7] Jangwon Park, “KoELECTRA: Pretrained ELECTRA Model for Korean”, <https://github.com/monologg/KoELECTRA>
- [8] <https://legalsearch.kr/>
- [9] 원준호, 조정진, 정상근, “기계독해를 이용한 판례 내 주요 정보 추출 방법”, 한국정보과학회 학술발표논문집, 1409-1411 (3 pages), 2019.
- [10] Ramshaw, Lance and Marcus, Mitch, “Text Chunking using Transformation-Based Learning”, arXiv preprint arXiv:cmp-lg/9505040.
- [11] Sandra Nemet and Dragan Kukolj, “Application of Data Mining in Patent Portfolio Technology Analysis”, Journal of Mechatronics, Automation and Identification Technology, Vol. 5, No.2, page 12-18, 2020.
- [12] 강슬기, 최윤수, 최성필, “생의학 분야 학술 논문에서의 개체명 인식 및 고나계 추출을 위한 언어 자원 수집 및 통합적 구조화 방안 연구”, 한국문헌정보학회지, 51(4), page 227-248, 2017.
- [13] Julien Rossi and Evangelos Kanoulas, “Legal Information Retrieval with Generalized Language Models”, Competition on Legal Information Extraction/Entailment(COLIEE), 4 pages, 2019.
- [14] Federico Barrios, Federico López, Luis Argerich, Rosa Wachenchauzer, “Variations of the Similarity Function of TextRank for Automated Summarization”, arXiv preprint arXiv:1602.03606, 2020.
- [15] Elena Leitner, Georg Rehm, Julian Moreno-Schneider, “A Dataset of German Legal Documents for Named Entity Recognition”, Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4478–4485, 2020.
- [16] 최정화, 박영택, “온톨로지 추론 기술 동향”, 정보과학회지 24(12), 47-55 (9 pages), 2006.
- [17] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman, "Glue: A multi-task benchmark and analysis platform for natural language understanding", arXiv preprint arXiv:1804.07461, 2018.