

툴민 논증구조 분석위한 딥러닝 기반 문장 경계 탐지 및 논증요소 분류에 관한 연구

정종진, 박노섭*

한국전자기술연구원, *,한림대학교 정보법과학연구소(교신)

mozzalt@keti.re.kr, *rspark@hallym.ac.kr

A Study on Deep Learning-based Sentence Boundary Detection and Argumentation Element Classification for Toulmin Argumentation Structure Analysis

Jung Jong Jin, Park Ro Seop

Korea Electronics Technology, Institute, *Hallym Univ.

요약

본 논문은 수사 중인 사건의 주요 쟁점에 대한 논증을 체계적으로 구성하여 사건을 객관적으로 고찰하기 위한 툴민(Toulmin) 논증 요소를 딥러닝 기법으로 자동 분류하는 방법을 제안한다. 툴민 논증 분석에 포함된 주요 요소들을 딥러닝으로 분류하기 위해서 자연어 처리 기법을 통해 필요한 수사문장들에서 툴민 요소에 맞게 문장을 자동 식별하고 분리하는 연구를 우선 제시하고, 분리된 문장들을 딥러닝 분류 기법을 통해 해당 문장이 어떤 툴민 요소인지 분류하는 연구를 제시한다. 실험 결과에서는 분리된 문장, 분류 모델의 모델 정확도 평가 결과를 제시하여 향후 경찰현장에서 범죄수사에 활용되는 수사결과보고서를 AI 분석 활용 하고자 할때 본 논문에서 제시한 모델이 활용될 가능성 또한 제시한다.

I. 서론

인공지능 기술이 많은 산업분야에 적용되기 시작하면서 법률/수사 영역에서도 자연어처리, 딥러닝 분석을 통한 수사 논증 분석등의 시도가 활발히 진행중이다. 오랜 전통 끝에 형사소송법, 검찰청법 개정으로 독자적 수사권, 경찰과 검찰의 상호협력관계 그리고 불송치 사건에 대한 수사 종결권을 확보하게 되었다. 경찰 입장에서는 권한이 커진 만큼 부담도 매우 늘어나게 된 셈이다. 이런 환경 변화는 경찰 수사의 결론에 대한 치밀한 검토가 요구된다는 점이다. 수사 종결권 확보와 책임 수사에 따라 수사오류를 최소화하고, 법정 증언에 대비하여 수사 단계 에서부터 변호인의 관점에서 수사 결과에 대한 검토가 필요하게 되었다는 의미는 수사 단계에서부터 수사 결론에 대한 검토를 도입해야 함을 의미한다. 경찰들은 사건을 종결할 때 공식적으로는 '수사 결과 보고서'라는 문건을 최종 작성하면서 관련 사건 수사를 마무리 하는데, 수사 종결권이 부여된 만큼 이 수사 결과 보고서는 어느 때 보다도 논리적으로 타당 해야되고, 주요 쟁점 즉 범죄가 성립하나, 아니냐 등의 주요 다툼에 대해 객관적 공정성을 갖춘 결론을 포함하고 있어야 한다.[1] 본 논문에서는 수사사건의 주요 쟁점에 대해 객관적 논증구조 분석을 위해 실용 논증법으로 알려진 툴민 논증법을 적용하여 쟁점에 대한 유죄와 무죄를 주장하는 양쪽의 논리를 문서에서 추출하고 분석하여 구조화 함으로써 쟁점에 대한 이슈를 논증학 적으로 오류가 없도록 필요한 요소를 딥러닝으로 분류하는 연구를 소개한다.

II. 관련연구

1. 툴민 논증 분석

툴민은 수사과정에서 발생하는 모든 논증이 고전적 논리의 구성요소(대전제, 소전제, 결론)로만 설명되지 않는다고 생각했고, 6가지 구성요소(datum(자료), claim(주장), warrant(정당화), qualifier(한정사), rebuttal

(반박), backing(보강))를 논증구조 분석의 요소로 그림1과 같이 제안하였다.[3]

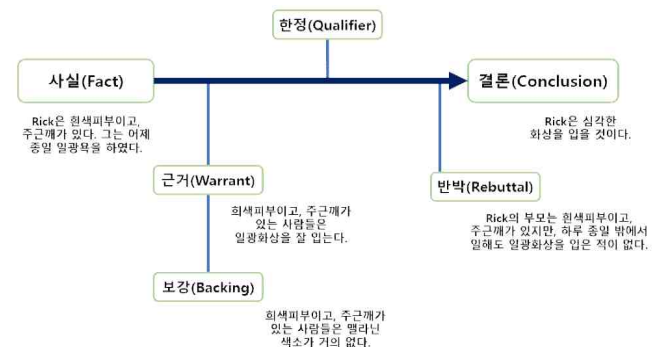


그림 1. 툴민이 제안한 법적 논증 구조화

2. 인공지능 기반 수사문서 분석

인공지능 기반 문장 자동 분류 시스템은 주어진 문장에 대해 자연어 처리, 인공지능 기반 분류 등을 통해 활용할 분류구조 즉, 클래스 중 하나로 분류하는 문제이다. 이 방법은 문장에 나타난 단어를 활용하고자 하는 서비스 영역에서 유의미한 핵심어휘 위주로 구성된 K 차원 벡터로 임베딩 시키고 n 개의 Convolutional Filter를 적용하여 문장의 의미를 분류 하는 방법이다. 문장 분류에서 전통적인 기계학습 방법으로는 나이브 베이즈(Naïve Bayes), 최대 엔트로피(Maximum Entropy), 서포트 벡터 머신(Support Vector Machines)이 있으며, 영화 리뷰 문서에 대해 긍정/부정 판단하는 이진 분류 실험에서는 서포트 벡터 머신이 가장 성능이 좋다.[3] 서포트 벡터 머신을 이용하여 다분류 문제를 풀 수 있는데 기본 아이디어는 이진 분류기를 여러 개 구성하는 방식으로 주로 활용된다.[4]

III. 톨민 논증 모델 설계

1. 톨민 논증 요소 정의

현대의 각종 수사 사건들의 범죄 수단이 정교함에 따라 수집된 증거에 대한 과학적 검증요소, 유사한 과거 사건에 대한 적용 수사논리 등이 중요해짐에 따라 그림 1에서 설명된 기본 톨민 논증 요소만으로는 논증 객관화하기 부족하여 한림대학교 정보법과학연구소에서 기존 톨민요소를 확장한 표1에 나타는 정보를 활용하여 분석한다.[5]

표 1 . 확장된 톨민 논증 요소 및 상세 설명

톨민 요소	설명	상세내용
주장(Claim)	변호인/피고인 및 판사의 주장 사실이나 증거	논증자가 달성하려고 하는 본 안건이며, 즉 논증자가 근거와 추론들을 바탕으로 주장하는 바이다.
자료(Datum)	경험적 지식, 법령	해당 범죄 사건에 대해 객관적으로 누구나 인지하고 있는 사실 또는 지식
정당화 (Warrant)	경험적 지식, 법령	규칙, 원칙, 추론 등의 명제이다. 자료, 보장들을 토대로 합리적으로 새롭게 판단 가능한 추론이며, 이 추론을 바탕으로 주장으로 연결됨
보장 (Backing)	참고 판례, 법 등	정당화에 대한 구체적인 근거이며, 증거, 증언, 공인된 감사 자료 등이 이에 해당됨
추론 (Inference)	합리적 판단	확보한 증거, 사실 기반 추론
전문가 의견 (Expert Option)	전문가 의견/증거	정당화된 주장을 패배 혹은 반박할 수 있는 예외 조건을 말할. 즉 하나의 사실로부터 서로 배치되는 증거로 다른 편의 추론, 있을수 있는 반대 경우들에 해당됨
쟁점결론 (Issue conclusion)	쟁점에 대한 최종 결론	해당 쟁점에 대해, 양쪽(피고/수사측)의 제시하는 증거, 추론, 주장들을 고려하여 판단한 최종 쟁점에 대한 결론

2. 톨민 논증 요소 자동 분석 모델 설계

본 논문에서 제안한 연구는 그림2의 개념도에서 설명하는 바와 같이 이미 작성된 수사결과보고서에 분석하여 도출한 사건 결론이 객관적으로 논란의 여지없이 잘 내렸는가를 검증하는데 도움을 주는 연구이다.[6]

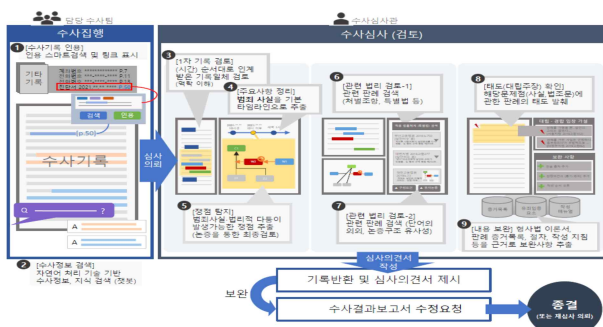


그림 2 . 톨민 논증 요소 시각화 개념 시나리오

이 과정은 입력된 수사문건에 대해 자연어처리, 텍스트분석, 딥러닝 분석을 통해 반자동으로 분석대상 문장이 구분되고, 구분된 문장이 각각 톨민의 어느요소에 해당 되는지 딥러닝 분류하여 그 결과를 톨민 논증모델에 맞춰 시각화를 통해 실현되며, 그 과정은 그림 3과 같다.

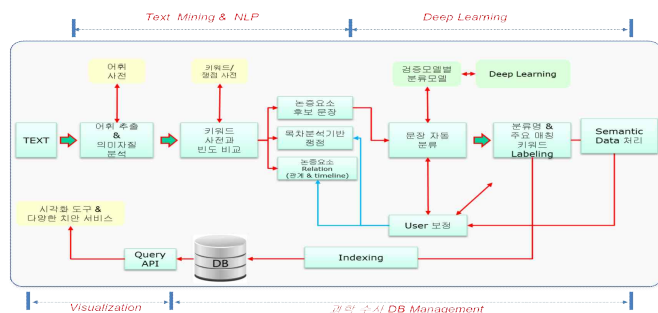


그림 3 . 톨민 논증 요소 분석 과정도

3. 톨민 논증 요소 분류를 위한 문장식별

톨민 논증 모델이 가진 요소들은 단어가 아닌 문장 단위로 구성된다. 따라서 입력된 문서에 포함된 모든 글들에 대해 범죄수사를 기록하는 문서 패턴에 맞춰 서로 의미를 달리하는 문장 단위로 구분해내야 한다. 여기서 말하는 문장단위는 마침표, 물음표 등 주요 구두점으로 끝나는 경우도 단위 문장이고, 접속사로 끝나는 경우도 분석단위 문장이 될 수 있다. 그림 4는 임의의 수사문건을 대상으로 분석단위 문장을 구분한 예이다.

그림 4 분석 단위 문장 구분 예시

위 그림4에 구분된 문장은 범죄유형마다 주로 사용되는 어휘, 접속사, 경찰관들의 오래된 문건작성 패턴등을 고려하여 구축한 문장구분 딥러닝 모델에 의해 자동 분리가 된다.

4. 딥러닝 기반 톨민 논증 요소 분류

그림 4의 추출된 문장들을 센텐스 단위 임베딩 하여 표1에서 정의된 톨민 요소들을 딥러닝 학습하고, 그림 5처럼 학습된 모델을 통해 새롭게 추출된 문장들을 자동분류하는 과정이다.

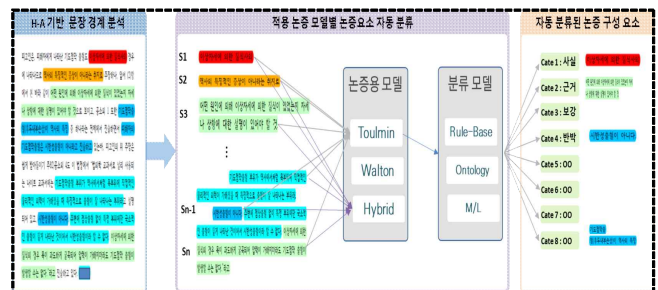


그림 5 Text 분석 기반 논증모델별 논증 구성요소 자동 분류 모델

우선 본 논문에서는 딥러닝 학습에 의존하지 않고, 간단한 사건이나 복잡도가 낮은 수사사건인 경우 사전에 정의된 Rule을 활용한 분류가 오히려 더 효율적인 경우가 있어 그림5에서처럼 분류방법을 규칙기반과 딥러닝 2가지 방식을 하이브리드 형태로 구성한다. 상기, 사전 지식 기반 논증 요소 분류 모듈은 논증 분석 전문가 지식을 활용할 수 있도록 DB(database)를 구축하고, 이를 실제 처리하기 위해서 논증 도메인에 특화된 정규표현식 패턴 집합을 개발하여 Python과 같은 각종 프로그래밍 언어에 쉽게 적용할 수 있도록 구성함. 이를 통해 전문가 지식을 활용하여 검증용 실제 데이터(Ground Truth Data) 확보가 가능하고, 기계학습 방식의 해석 가능성 문제를 일부분 해소한다. 딥러닝 접근법에 의한 논증 요소 분류 모듈은 기본적으로 톨민 논증구조를 따르고, 이를 위한 메타언어를 적용하여 처리한다. 논증 요소를 추출하고 검증 및 정제함에 있어서 BERT나 GPT 계열과 같은 문장내 단어 사이의 관계 및 컨텍스트 정보 이용이 가능한 초거대 인공지능경량 기법을 활용하여 분류 및 일부 정보를 생성하는 처리를 수행하는 것을 지향하고 분류 결과의 성능 평가를 위한 데이터 파이프라인 구축이 필요하다.

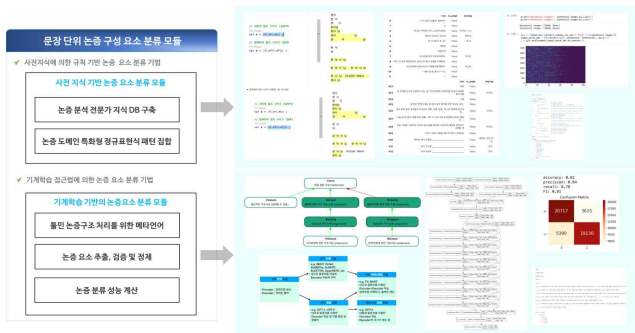


그림 6 툴민 논증구성요소 분류 모델 및 실험 결과

IV. 실험결과

4-1. 실험 구성

4장에서는 수사문건 중 해당 사건의 주요 쟁점과 쟁점을 토대로 피고인과 수사기관의 논리가 작성된 부분을 자동 추출하고, 각 문단마다 개별 툴민 논증요소로 예측되는 문장을 자동 분리하여 해당 문장이 툴민 요소중 어떤 것에 해당되는지를 자동 분류하는 시스템을 설계하고 이 기능을 수행하는 모델에 대한 정확도를 결과로 제시한다. 실험을 위해 수사문건은 살인사건 법원 1심 판결문을 1500건을 활용한다. 1500건의 1심 살인사건 판결문을 대상으로 다수의 현직 수사관들이 직접 태깅을 실시한 실험 데이터셋을 <표 2>와 같이 구성한다.

표 2. 실험데이터 구성

요소	설명			
실험 문서	2000년 이후 살인사건 1심 판결문 1500건			
구성요 소별 총 실험셋	요소	실험셋 규모	학습셋(80%)	평가셋(20%)
	주장(Claim)	7,945	6,356	1,589
	자료(Datum)	84,265	67,412	16,853
	정당화 (Warrant)	5,328	4,262	1,066
	보강 (Backing)	3,457	2,766	691
	추론 (Inference)	2,273	1,818	455
	전문가 의견(Expert Opinion)	2,162	1,730	432
	쟁점결론(issue conclusion)	2,103	1,682	421
	합계	107,533	86,026	21,507

범죄사실 문장들을 대상으로 자연어 분석 결과 추출된 요소들(단어 및 문장들)을 시각화에 표현하기 위한 스토리라인 구성요소로 분류하는 분류 모델의 성능을 평가하기 위해 우선 Confusion Matrix로 결과를 도출한다. 이를 바탕으로 분류 모델의 객관적 성능, 개선점 등을 도출하기 위해 ① 분류 모델이 Positive로 분류한 것 중 실제 정답인 비율을 점검하기 위해 Precision(정밀도)을, ② 실제값이 정답인 것중 분류 모델이 Positive로 분류한 비율을 점검하기 위해 Recall(재현율)을, ③분류 모델이 바르게 분류한 비율을 점검하기 위해 Accuracy를, ④정밀도와 재현율의 조화평균을 점검하기 위한 F1-Score를 같이 도출된다

4-2. 모델 정확도 결과

문장 경계 탐지는 툴민 논증 요소의 대상이기 때문에 표2에서 설명된 각 요소별 총합은 분리된 문장수 이다. 따라서 본 논문에서 제안한 문장경계 탐지 모델의 실험 셋 107,533 문장중 80%에 해당하는 훈련셋으로 학습하여 20%의 평가셋으로 측정된 모델의 정확도는 표 3과 같다.

모델은 BERT를 파인튜닝한 Sentence-BERT를 활용하여 학습을 하였고, 최종 분류에서는 다대다 분류에 해당하여 Softmax 분류를 진행하였다.

표 3. 툴민 논증 요소 분류 모델 성능 평가

요소	실험셋	Precision	Recall	Accuracy	F1-Score
주장(Claim)	7,945	0.89	0.81	0.91	0.84
자료(Datum)	84,265	0.75	0.73	0.60	0.74
정당화 (Warrant)	5,328	0.88	0.81	0.75	0.84
보강 (Backing)	3,457	0.89	0.80	0.76	0.85
추론 (Inference)	2,273	0.93	0.88	0.81	0.87
전문가 의견(Expert Opinion)	2,162	0.90	0.84	0.79	0.86
쟁점결론(issue conclusion)	2,103	0.97	0.94	0.89	0.91

표 3의 결과를 살펴보면 전체적으로 모든 논증요소들이 70% 이상의 정확도로 분류가 됨을 확인 할 수 있었고, 요소별로는 주장이나 쟁점결론 분류가 정확도가 높다. 이는 주장이나 쟁점결론의 문장 길이가 비교적 길지 않고 문건 내에 사용되는 단어나 어구등의 특징이 명확하여 모델 분류 정확도가 높은 반면, 자료(Datum)인 경우 사실, 증거 등등 다양한 사실을 표현하는 문장이다 보니 문장 길이와 사용되는 어휘가 다양하여 분류 정확도가 다른 요소에 비해 낮다. 이는 개체명 인식의 정확도를 높혀 분류 정확도를 개선해야 한다는 향후 연구 방향을 제시하고 있다.

V. 결론

본 논문에서는 현직 수사관들이 접수된 범죄 사건에 대해 올바른 수사 방향을 수립하거나, 이미 종결된 사건 보고서의 주요 쟁점등에 대해 논리적 모순 없이 수사가 이루어 졌는지, 양쪽의 주장과 의견을 균형있게 잘 고려했는지를 검토 할 수 있는 툴민 논증요소 추출 연구를 살펴보았다. 본 연구는 많은 사건을 살펴봐야 하는 수사관들에게 손쉽고 빠르게 쟁점을 살펴 피의 객관적인 사건수사,검토를 할 수 있도록 도움을 주는 연구이다.

ACKNOWLEDGMENT

이 논문은 2022년도 정부(경찰청)의 재원으로 지원받아 수행된 연구결과임 [내역사업명: AI 기반 범죄수사 지원 / 연구개발과제번호: PR10-02-000-21]

참 고 문 헌

- [1] 윤원대, “디지털 증거 수집절차 개선방안에 관한 연구”, 디지털포렌식 연구 제9권 제2호, pp. 40-41, 2015.
- [2] 강남화, 이은경 “논변,논의 그리고 논증:개념의 명료화를 위한 문헌조사 연구”, 한국과학교육학회지, 33권 6호, pp. 1119-1138, 2013
- [3] B. Pang, L. Lee, S. Vaithyanathan, “thumbs up?Sentiment Classification using Machine LearningTechniques”, in Proceedings of the Conference onEmpirical Methods in Natural Language Processing,pp. 79 - 86, 2002.
- [4] M. Aly, “Survey on multiclass classification methods”, Neural Networks, (11):1 - 9, 2005.
- [5] 한림대학교, 정보법과학연구소, “논증구조 수사데이터분석 가이드라인 V2.0”, 2022
- [6] 한림대학교, 정보법과학연구소, “AI과학수사기술개발 1차년도 결과보고서”,2021