수사문서 대상 논증 분석용 학습데이터 증식 방법 연구

정종진, 박종빈*
한국전자기술연구원
mozzalt@keti.re.kr, *jpark@keti.re.kr

A study on how to augment training data for argument analysis of investigative documents

Jung Jong Jin, Park Jong Bin*
Korea Electronics Technology Institute

요 약

본 논문은 형사재판 판결문, 수사결과보고서 등에 포함된 주요 범죄 관련 내용을 인공지능 기법으로 분석하는데 필요한 학습데이터를 확장하기위한 방법 으로서 유의어 사전을 통한 증식 방법과 증식된 데이터의 유효성 검증 결과를 제시한다. 우선 학습데이터를 증식하기 위한 주요 수단으로 공개된 한국어 말뭉치와 범죄수사 도메인에서 미리 정의된 유의어 어휘사전을 활용하여 유의어 교체, 반대어교체 등 다양한 방식으로 원본 문장이 가진 의미를 유지하되 서로 다른 형태로 증식하는 방법을 소개한다. 이런 방식으로 증식된 데이터들이 원본 데이터와 유사한지를 판단하는 방법을 소개한다. 결국 본 논문에서 소개된 증식 방법을 통해 확장된 학습 데이터가 분석 모델의 정확도 향상 시키는 주요한 수단으로 활용 될 수 있다

Ⅰ. 서 론

최근 LegalTech(법률서비스 분야에서 IT 기술을 이용한 분석기법)가 관심을 받으면서 판결문, 법률문서 등을 자연어처리, 딥러닝 방법 등을 활용하여 자동 법률해석, 법률조언, 범죄 분석등의 서비스가 등장하고 있다. 실제로 이런 기술과 서비스들은 범죄 미래에 발생할 수도 있는 범죄를 예방하거나 현재 수사 중인 사건을 해결함에 있어 매우 유용한 도구이다. 이를 AI, 특히나 딥러닝을 활용하여 분석을 하고자 할 때는 법률문서로부터 구축한 학습데이터의 양과 질은 LegalTech 서비스 성능을 결정하는 매우 중요한 요소이다.[1] 현실적으로 이런 학습데이터를 구축하는 단계는 통상 법률서비스에 종사하는 전문가 (해당 용어와 문장의 의미가 무엇인지를 알 수 있는 전문가) 또는 교육을 받은 사람들에 의해 구축하곤 하는데, 딥러닝에 필요한 학습데이터량이 매우 많다는 점을 감안하면 사람이 직접 모두 다 하기에는 시간과 비용이 매우 소비되는 어려움이 존재한다. 이런 이유로 통상 LegalTech 서비스 준비중 업체들도 필요한 양질의 학습데이터를 확보하기 어려워 서비스 개발이 매우 어려운점이 있는 것도 업계의 현실이기도 하다. 따라서 본 논문에서는, 법률 서비스 도메인 전문가가 태깅(Tagging)한 학습데이터를 기반으로, 기본적인 패턴을 발견한 후 법률분야 유의어/반대어 사전을 활용하여 증식을 한 후 이를 학습데이터로 재활용 함으로써, 보다 빠르고 효과적으로 인공지능 분석에 학습데이터를 확보하고, 편향성이 발견되는 경우 데이터를 보정 할 수 있는 방법을 제공하는 기술을 제안한다.

Ⅱ. 논증용 학습데이터 증식 방법

1. 개념

서론에서 밝혔듯이 AI 활용한 수사결과 분석을 하기 위해 반드시 필요한 학습데이터를 구축함에 있어 통상 해당 분야 전문가들이 직접 태깅하여 학습데이터를 확보한다. 경우에 따라서는 학습데이터를 보다 쉽고 빠르게 하기 위해 태깅툴(Annotation 툴)을 활용하여 자동 태깅(Auto Labeling) 할 수 있는데, Auto 태깅을 위해서는 역시 학습데이터로부터 학습된 모델을 활용해야 하므로 결국 양질의 학습데이터가 필요한 건 마찬가지 이다. 따라서 일정 규모로 전문가에 의해 확보된 학습데이터와 해당분야 유의어/반대의 사전을 기초 데이터로 활용하고, 증식을 통해 보다 쉽고 빠르고 효과적으로 학습데이터를 확장/생성 방법이 필요하여 그림 1에서와 같이 이미 확보된 학습용 문장에 유의/반대어를 활용하여 5가지 증식 방식을 통해 증식된 학습 문장을 생성한다.

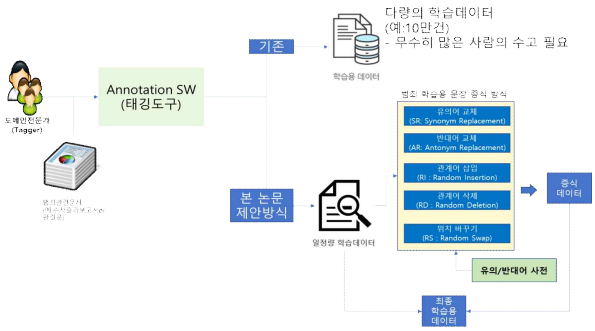


그림 1. 증식을 통한 분석용 학습데이터 증식 방식 과정 개념

2. 유의어/반대어 사전

유의어 반대어 사전은 해당 서비스 도메인에서 주도 활용되는 어휘들간 유사한 의미와 반대되는 의미를 가진 단어들의 집합이다. 통상 데이터 분석 영역에서 흔히들 말하는 어휘사전과 같은 개념이다. 유의어/반대어 사전이 필요한 이유는 같은 의미라도 문서를 작성하는 사람들의 문장 쓰는 스타일에 따라 조금은 다른 단어, 다른 조사, 어구를 사용하기 때문이다.[2] 그림2는 본 논문에 활용된 유의어/반대어 사전의 예이다. 유의어/반대어 사전은 증식을 하려는 목적에 맞춰 미리 준비된 어휘 사전을 활용한다.

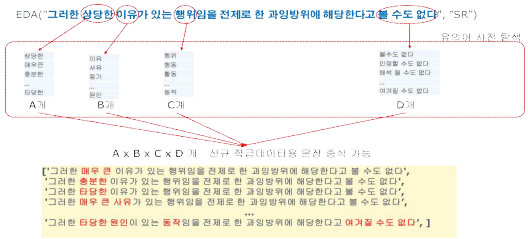
단어	종류	유의어	반대어
종결	명사	완결	착수
		완료	시작
		마무리	
~한것 으로 보아	구문	~ 한것으로 판단한 결과	

그림 2. 유의어/반대어 사전 예

3. 학습용 문장 증식 방법

도메인 전문가 집단을 통해 생성된 학습데이터가 일정 정도 (증식을 하기에 충분한 규모) 구축이 되면, 그림 1에서 설명된 바와 같이 5가지 방식, 즉, 유의어교체, 반대어교체, 관계어 삽입, 관계어 삭제, 위치 바꾸기이며 각 방식에 대한 상세 설명은 표1과 같다.[3][4]

표 1 증식 유형별 상세 설명 및 그 과정

증식방식	설명 및 과정
유의어 교체 (SR: Synonym Replacement)	• 준비된 학습데이터용 문장 중 품사/어휘 분석/개체명 인식을 통해 주요 명사/어구를 식별한다. • 식별된 명사/어구 등과 유의어(뜻은 같고 다른 표현)으로 자동 교체하여 다수의 문장을 만든다. • 한 문장에 다수의 교체 가능 유의어 들이 나온다면 전체 문맥의 의미를 훼손하지 않는 조건하에 교체된 유의어로 재표현된 문장들을 생성해낸다 (증식)  • 증식된 문장들은 유효성 검증을 통해 문맥 훼손에 대한 검증을 거친다.
반대어 교체 (AR: Antonym Replacement)	• 서술어 위주로 반대말로 교체하는 방식으로, 반대 뜻을 가진 문장을 만들어 내는 과정이다. • 유의어 교체와 병행하여 활용되나, 반드시 반대 서술어를 사용해야 한다는 점에서 조금의 차이가 있다. • 위 그림에에서 A x B x C x D' 의 문장수를 만들어 낼 수 있다 (D'은 D의 반대어 표현 가능 수)
관계어 삽입 (RI : Random Insertion)	• 관계어 삽입은 형용사와 부사를 확보된 문장에 삽입함으로써 학습용 문장을 증식하는 방식이다. • 삽입 가능한 형용사와 부사 목록은 유의어/반대어 사전에 이미 포함되어져 있다. • 즉 “법률분석”용으로 미리 분석하여 정리/확보한 사전 지식의 일종이다. 원문: 그러한 상당한 이유가 있는 행위임을 전제로 한 과잉방위에 해당한다고 볼 수도 없다 증식문장 예시: 1. 그러한 상당한 이유가 있는 행위임을 전제로한 과잉방위에 해당한다고 *충분하게* 볼 수도 없다 2. 그러한 상당한 이유가 있는 행위임을 전제로한 과잉방위에 해당한다고 *확실하게* 볼 수도 없다 3. 등등
관계어 삭제 (RD : Random Deletion)	• 관계어 삽입과 반대의 경우 • 문장에 존재하는 형용사/부사를 삭제함으로써 증식된 문장 생성 원문: 그러한 상당한 이유가 있는 행위임을 전제로 한 과잉방위에 해당한다고 볼 수도 없다 증식문장: 1. 그러한 ~~상당한~~ 이유가 있는 행위임을 전제로한 과잉방위에 해당한다고 볼 수도 없다 2. 기타 등등
위치바꾸기 (RS : Random Swap)	• 열거형, 사실이 포함된 문장에서 동일 수준의 단어들을 위치를 바꿈으로써 문장 생성 원문: 피고인은 17일 오후 마트에 들러, 테이프, 큰비닐봉지, 커터칼 등을 구매하여 숙소로 빠르게 이동하였다. 증식문장: 1. 피고인은 17일 오후 마트에 들러, 큰비닐봉지, 테이프, 커터칼 등을 구매하여 숙소로 빠르게 이동하였다. 2. 기타 등등 • 목적어 - 부사 순으로 표현된 문장일부를 부사-목적어 순으로 교체하여 문장 생성 원문: 피고인은 17일 오후 마트에 들러, 테이프, 큰비닐봉지, 커터칼 등을 구매하여 숙소로 빠르게 이동하였다. 증식문장: 1. 피고인은 17일 오후 마트에 들러, 큰비닐봉지, 테이프, 커터칼 등을 구매하여 빠르게 숙소로 이동하였다. 2. 피고인은 17일 오후 마트에 들러, 빠르게 큰비닐봉지, 테이프, 커터칼 등을 구매하여 숙소로 이동하였다.

3. 학습용 문장 증식 문장 생성

표 1에 설명된 증식 방법별로 얼마의 비중으로 증식할지에 대해 가중치를 부여하여 증식 과정을 식 (1)과 같이 진행된다.

$$f_{augment}(P_{sr}, P_{ar}, P_{ri}, P_{rd}, P_{rs}) \quad (1)$$

단 $P_{sr} + P_{ar} + P_{ri} + P_{rd} + P_{rs} = 1$

Ⅲ. 실험결과

1. 증식 수행

식(1)에 의해 판결문 내 포함된 원문 대상으로 유의어교체를 50%, 반대어교체를 10%, 랜덤 삽입 20%, 랜덤 삭제 10%, 위치교체 10% 비율로 15개 문장을 증식/생성 실험하였다.

```
EDA("따라서 피고인 및 변호인의 과잉방위 주장은 받아들이지 않는다", param_sr=0.5,
param_ar=0.1, param_ri=0.2, param_rs=0.1, p_rd=0.1, num_aug=15)
```

실험 결과 생성된 증식 문장들은 그림 2와 같다.

```
python
['따라서 피고인 및 변호인의 주장은 과잉방위 받아들이지 않는다',
 '따라서 피고인 및 변호인의 과잉방위 주장은 받아들이지 않는다',
 '따라서 피고인 및 주장은 과잉방위 변호인의 받아들이지 않는다',
 '따라서 과잉방위 주장은 받아들이지 않는다',
 '따라서 피고인 및 변호인의 과잉방위 주장은 받아들이지 않는다',
 '따라서 피고인 및 변호인의 과잉방위 주장은 받아들이지 않는다',
 '따라서 피고인 및 변호인의 과잉방위 주장은 받아들이지 않는다',
 '따라서 피고인 및 변호인의 주장은 과잉방위 받아들이지 않는다',
 '따라서 피고인 및 변호인의 과잉방위 주장은',
 '따라서 피고인 및 변호인의 과잉방위 주장은 받아들이지 않는다',
 '변호인의 피고인 및 따라서 과잉방위 주장은 받아들이지 않는다',
 '따라서 피고인 및 변호인의 과잉방위 주장은 받아들이지 않는다',
 '따라서 피고인 및 변호인의 과잉방위 주장은 받아들이지 않는다',
 '따라서 피고인 및 변호인의 과잉방위 주장은 받아들이지 않는다',
 '따라서 피고인 및 변호인의 과잉방위 주장은 받아들이지 않는다']
```

그림 2. 증식유형 별 가중치에 따른 증식문장 생성 결과

위 증식 실험에 포함되는 파라미터들의 비중 조절을 통해, 현재 확보된 학습데이터의 분포가 편향되어 있다던가, 분석에 필요한 데이터들의 확보 상태가 상대적으로 부족한 경우 위 파라미터를 조정함으로써 원하는 데이터를 확보 가능하다.[5]

2. 증식 데이터의 유효성 검증

증식으로 인해 생성된 데이터가 분석 결과를 더 나쁘게 만들면 안되기 때문에, 증식된 문장들이 원본데이터와 유사한지, 그리고 증식된 문장들을 확대하여 모델의 정확도는 개선되는지를 그림 3과 같이 검증한다.

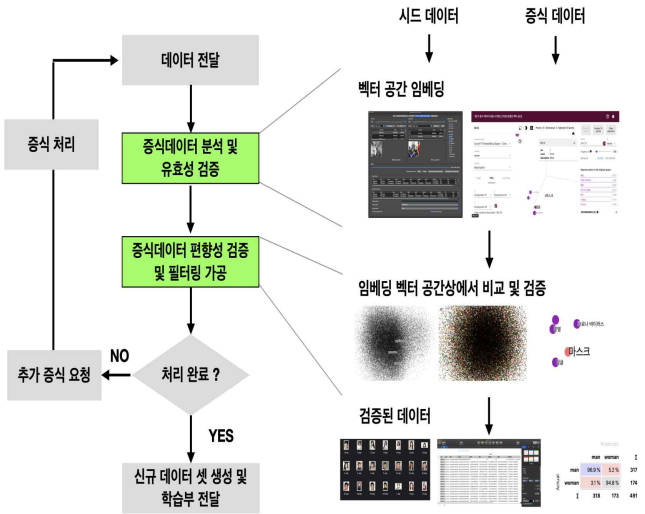


그림 3. 증식된 결과 유효성 검증 및 결과에 따른 편향성 보정 과정

각 문장들은 문맥과 의미를 내포하고 있기 때문에 단순한 단어 유사도만을 측정하는 방식이 아닌 고차원 벡터공간에서 임베딩 하여 클러스터링이 되는지를 확인하고 경계값 이상이 되는 이상치를 배제하여 유효성을 검증한다. 유효성 검증 결과에 대한 시각화 표현은 그림 4와 같다.

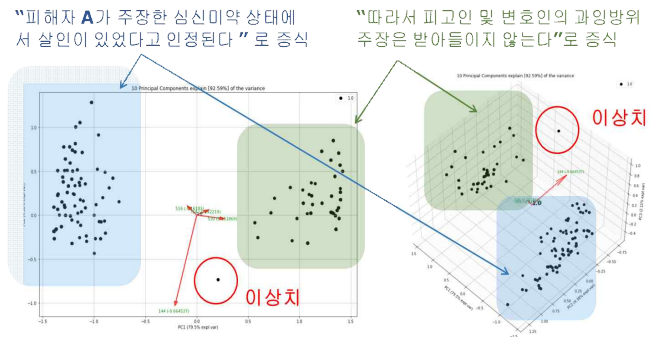


그림 4. 증식된 문장대상 유효성 결과 시각화 표현

Ⅲ. 결론

본 논문에서는 범죄수사 과정에서 발생한 문서에서 중요한 정보를 추출하여 분석하는데 필요한 학습 데이터를 증식을 통해 빠르고 효과적으로 확보 할 수 있는 방법을 제안하였다. 기존 해당 도메인 전문가에 의해 수동으로 구축한 일정규모의 학습데이터를 바탕으로 유의어/반대어 사전을 활용하여 증식함으로써 증식된 데이터의 정확성 향상에 도움이 됨을 확인하였고 또한 증식 과정에서 발생 가능한 학습데이터 편향성을 보다 효과적으로 보완하는 방법으로 활용 가능함을 확인하였다. 향후 연구에서는 적용 가능 죄종을 확대하고 각 죄종별 유의어/반대어 어휘 사전을 확장하여 보다 많은종류의 수사문건에 활용 될 수 있는 연구를 진행하고자 한다.

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] 정종진, 박종빈, 박성주 “주요 범죄사실 개요 분석을 통한 범죄사실 타임라인 자동 작성 시스템 연구”, ACK 2021 학술발표대회 논문집, 28권 2호, pp. 622-625, 2021.

[2] 윤원대, “디지털 증거 수집절차 개선방안에 관한 연구”, 디지털포렌식 연구 제9권 제2호, pp. 40-41, 2015.

[3] 홍승표,임선영,고으뜸,홍성초,김정운,임중연, “판결문 분석을 통한 범죄 빅 데이터의 활용 가능성에 관한 소고”, 한국경찰학회보, 22권 6호, pp.209-230, 2020.

[4] 정종진, 박종빈, “증식된 데이터 대상 유효성 검증 방법 연구”, 한국멀미디어학회 2022년도 한국멀티미디어학회 춘계학술발표대회 논문집 제25권 1호, 2022

[5] 박시현, “판결문의 특성”, 「열린정신 인문학연구」, 8권, pp. 179-194, 2007