

사전학습 기반의 법률문서 요약 방법 비교연구

김의순*, 임희석**

*고려대학교 정보통신대학원 빅데이터융합학과

**고려대학교 컴퓨터학과

kes1492@korea.ac.kr, limhseok@korea.ac.kr

Comparative study of legal document summary method based on pre-trained model

EuiSoon Kim*, HeuiSeok Lim**

*Graduate School of Computer&information Technology, Korea University

**Dept. of Computer Science and Engineering, College of Informatics, Korea University

요 약

법률 문서는 일반 사용자가 이해하기 어려운 용어로 이루어져 있고 특히 장문의 문서가 많아 법률시스템에 종사하는 종사자들 또한 많은 양의 문서를 읽기가 어려운 현실이다. 이에 문서 요약 방법중 딥러닝 기반의 사전학습 모델을 적용한 추출요약기반, 생성요약 방법론과 딥러닝 이전의 핵심문장 추출 방법론을 비교하여 법률용어의 요약성능에 대한 비교 평가를 수행하고자 하며 추후 연구과제로 법률문서에 특화된 요약 모델을 만들어보고자 한다.

1. 서론

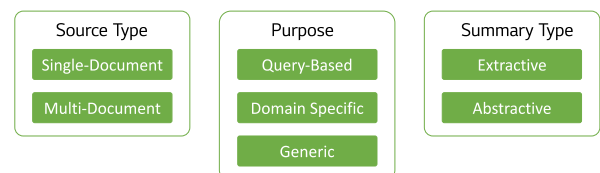
법률 인공지능은 판결문, 계약서, 법적자문등 텍스트문서에 기반한 TASK에 인공지능을 적용하기 위한 연구이다. 법률문서는 비 전문가에게는 생소한 용어로 이루어져있고 법조인들에게도 방대한양의 문서로 인해 법률문서생성, 법률 엔티티 추출, 법률 QA, 법률문서요약등의 TASK를 수행하는데 많은 시간이 소요되는 일이다[1]. 이에 설명가능한 인공지능을 이용한 법률문서 유사도 분석 서비스에 대한 연구[2]와 텍스트 마이닝 기법을 적용한 법률정보 시스템에 대한 연구가 있었다.[3]

법률 문서역시 텍스트 기반으로 이루어져 있으므로 텍스트를 요약하는 방법론중 딥러닝 기반의 방법론을 활용하여 각각 모델의 성능을 비교 분석하여 법률문장의 모델에 대한 평가를 수행하고자 하며 추후 연구과제로 법률문서의 요약에 특화된 모델을 구현해보고자 한다.

2. 관련연구

텍스트로 이루어진 문서들을 요약하는 행위를 summarization 이라 하며, 문서의 분류종류에 따라 추출요약방법(Extractive approaches)와 생성요약방법(Abstractive approaches)로 나뉘며 입력 문장의 종류에 따라 Single/Multi Document, 요약의 목적에 따라

Query-based, Domain specific, Generic 한 방법으로 나뉘 수 있으며[4], 본 논문에서는 Summary Type 의 두가지 추출요약과 생성요약 방법에 대한 비교 연구를 수행하고자 한다.

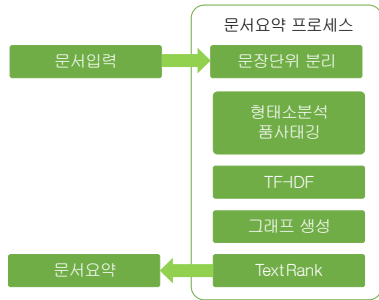


(그림 1) 문서요약 분류

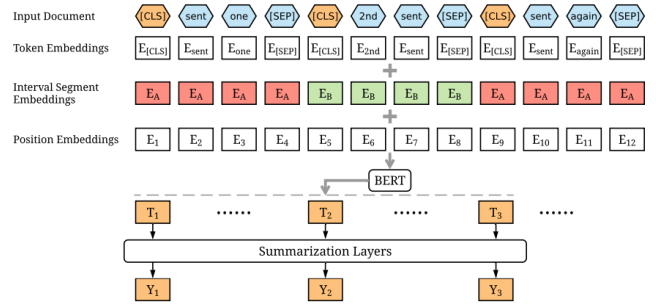
2-1. 딥러닝이전 추출모델

딥러닝 이전의 요약방법은 추출요약 기반의 주어진 문서의 집합들 안에서 핵심 단어들이나 문장들을 선택하는 방법으로 주어진 문장들과 단어들 안에서의 요약을 하기때문에 표현이 제한 된다는 단점이 있고 딥러닝 이전의 추출요약 방법에는 대표적으로 TextRank[5]가 많이 사용되고 있다.

TextRank 는 키워드 추출기능과 핵심 문장 추출 기능 두가지를 제공하며 단어 그래프 기반으로 키워드와 핵심문장을 선택하여 문서집합을 요약하는 방법이며, 법률 문서에서도 베이지안 최적화 기반 접근방법에 기반한 TextRank 최적화 방법이 제안되었다.[6]



(그림 2) TextRank 프로세스



(그림 3) BertSum 구조도

2-2. 딥러닝기반 모델

딥러닝 기반의 추출요약은 사람이 직접 요약하는 것처럼, 문서의 내용을 바탕으로 요약문장을 만드는 방법이다. Attention mechanism은 Sequence-to-sequence 모델은 디코딩(decoding) 과정에서 각 시간별 출력을 예측할 때 입력 열에서 해당 출력을 결정하는데 필요하다고 판단되는 입력 시퀀스의 hidden state를 참조하여 사용하는 방법이며,[7]. Copy mechanism은 Sequence-to-sequence 모델의 디코딩 과정에서 문장을 생성할 때 필요한 어휘가 출력 사전에 없는 문제(Out-of-Vocabulary)와 고유명사들의 출력 확률이 작아지는 문제를 해결하기 위해 고안된 방법으로, 출력에 필요한 어휘를 입력 열에서 찾아 출력 없이 복사(copy)하는 사용하는 방법이다[8]. 본 논문에서는 pre-trained BERT 기반으로 구현된 여러 sentence를 하나의 인풋으로 사용하고 interval segment embedding을 추가한 변형 인풋 embedding을 사용한 BertSum[9]모델을 추출모델로, Sequence-to-sequence 형태로 denoising 기법을 활용한 BART[10] 모델을 생성모델로하여 비교 실험을 진행하였다.

3. 모델설명

3.1.TextRank

TextRank [5]는 word graph 나 sentence graph 를 구축한 뒤, Graph ranking 알고리즘인 PageRank 를 이용하여 각각 키워드추출과 핵심 문장추출이 가능한 모델로 토크나이저는 코모란을 사용하였다.

3.2.BertSum

딥러닝 추출모델로 Pre-trained BERT 로 Hugging face transformers 라이브러리를 사용한 KoBertSum 모델을 사용하였다. KoBertSum 은 BertSum[9]모델을 기반으로 하는데 BertSum 은 Bert 위에 Inter-sentence transformer 2-layers 를 추가하고 fine-tuning 하여 추출요약 모델로 구성되어있다.

모델의 파라미터수는 103M 개이며 Bert 의 embedding 은 word embedding 은 8002 개, position embedding 은 512 개, Token embedding 은 2 개로 이루어져있고 batch size 는 3000, drop-out 은 0.1 로 설정하여 학습을 진행하였다.

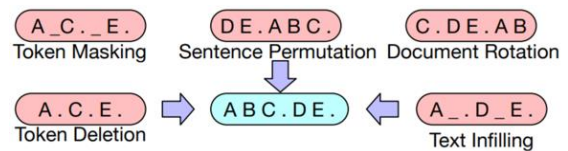
<표 2> KoBertSum 모델 Parameter

KoBertSum Model	Value
Pre-trained model	monolog kobert
Parameter 개수	103M
Drop out	0.1
Batch size	3000
Word Embedding	8002
Position Embedding	512
Token type embedding	2

3.3.BART

BART(Bidirectional and Auto-Regressive Transformers) [10]는 페이스북(Facebook)에서 발표한 Pre-training Sequence-to-Sequence Transformer 구조를 사용한 모델로 Sequence-to-sequence 구조로 만들어진 denoising auto-encoder 기반으로 손상된 텍스트를 복구하도록 모델을 학습시키는 방법이다.

본 논문에서는 Text infilling 노이즈 함수로 40G 이상의 한국어 텍스트에 대하여 학습한 SKT-AI 팀이 한국어에 맞게 수정한 KoBART 모델을 사용하여 생성모델로 사용하였다.



(그림 4) BART 학습방법

모델의 파라미터수는 135M 개이며 Base model 로 KoBART base 모델을 기반으로 법률, 뉴스 데이터로 각각 학습하여 실험을 진행하였다.

<표 3> KoBART 모델 Parameter

KoBART Model	Value
Pre-trained model	KoBART base
Parameter 개수	124M
Drop out	0.1
Batch size	4
Position Embedding	1026

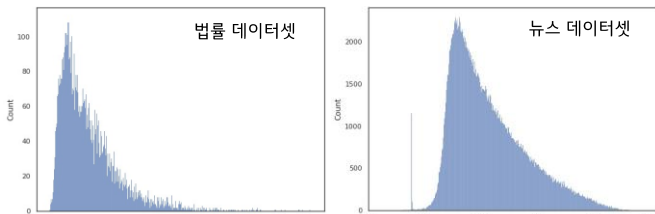
4. 실험 및 결과

4.1. 실험 데이터

본 논문에 사용된 데이터는 AI HUB 에 공개된 한국어 ‘문서요약 텍스트’로 원본텍스트와 요약 텍스트의 쌍으로 되어있는 법률데이터 16,695 건, 뉴스데이터 270,697 건을 사용하였다.

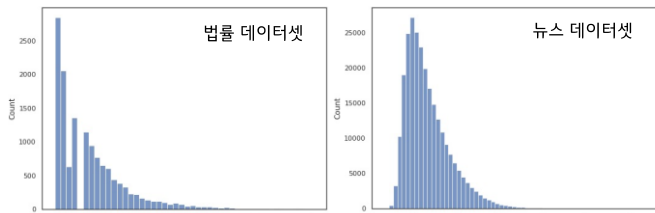
<표 4> Dataset 종류 및 개수

Data 종류	Train	Valid	Test	Total
법률 데이터	13,000	1,000	2,695	16,695
뉴스 데이터	247,662	13,035	10,000	270,697



(그림 5) 데이터셋 단어분포

데이터셋의 문장내 단어 분포비율을 비교해보면, 법률 데이터가 평균 44 개(min 7 ~ max 210)이며, 뉴스 데이터셋은 평균 213 개(min 23~ max 488)로 이루어져 있는것으로 분석되었으며 이는 뉴스문장이 법률문장에 비해 많은 단어로 구성되어 있음을 알 수 있다.

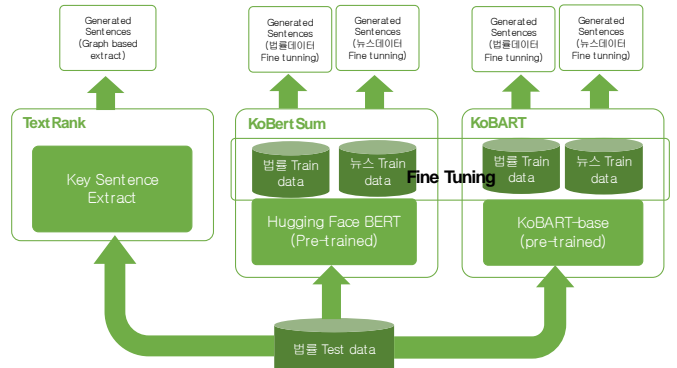


(그림 6) 데이터셋 문장분포

데이터셋의 문장 분포 비율을 비교해 보면, 법률 데이터셋은 평균 9.9 개(min 4 ~ max 50)이며, 뉴스 데이터셋은 평균 12 개(min 3 ~ max 64)로 이루어져 있는것으로 분석 되었고 뉴스 문서의 문장개수가 법률문서에 비해 많은 문장으로 구성되어 있음을 알 수 있다.

4.2. 실험방법

실험은 3 가지 모델 1) TextRank 2)KoBertSum 3)KoBART 로 진행하였고 사전학습 기반 모델인 KoBertSum 과 KoBART 는 각각 법률/뉴스 학습 데이터로 fine tuning 후 법률 테스트 데이터로 Text 생성을 진행하였다.



(그림 7) 모델별 실험방법

4.3. 평가방법

각 모델에 대한 성능 평가 방법으로 텍스트 요약 타스크에서 흔히 사용하는 ROUGE[11] (Recall-Oriented Understudy for Gisting Evaluation) 를 사용한다. ROUGE 는 텍스트 자동 요약, 기계 번역 등 자연어 생성 모델의 성능을 평가하기 위한 지표이며, 모델이 생성한 요약본 혹은 번역본을 사람이 미리 만들어 놓은 참조본과 대조해 성능 점수를 계산하는 방법을 사용한다.

본 논문에서는 AI HUB 에 공개된 데이터의 original text 와 reference text 를 활용하여 각 모델에서 생성된 text 와의 ROUGE F-1 score 를 바탕으로 비교 실험 하였다.

4.4. 실험결과

TextRank 가 Rouge-1 의 성능이 다른 모델에 비해 떨어지지 않는것으로 측정되었는데 이는 TextRank 에 문장 입력을 넣을때 얼마나 문장끼리의 분절을 많이 하느냐에 따라 성능이 달라진것으로 예측되며, 전체적으로는 KoBART 모델에 24 만건의 뉴스데이터로 Fine tuning 한 모델의 성능이 제일 우수한것으로 측정되었다.

KoBART 모델은 Base 모델의 성능도 다른 모델에 비해 우수한것으로 측정되었으며 BetSum 기반의 모델은 의외로 성능이 떨어지는 것으로 측정되었다.

<표 5> 모델별 요약 성능 비교

Models	Rouge-1	Rouge-2	Rouge-L
TextRank	0.542	0.39	0.276
KoBertSum - Legal	0.484	0.33	0.334
KoBertSum - News Large	0.472	0.36	0.366
KoBART - Base	0.534	0.42	0.446
KoBART - Legal	0.544	0.44	0.464
KoBART - News Large	0.57	0.47	0.508

이는 문서 생성에서 사전학습 모델이 얼마나 많은 Corpus 로 학습 되었는지에 따라 도메인에 관계없이 우수한 성능을 보이는 것으로 유추해 볼 수 있으며 BART 모델의 Fine tuning 결과에서 법률 데이터 14,000 건은 뉴스 데이터 24 만건에 비해 상대적으로 작은 데이터셋으로 학습되었고, 특히 사전학습 모델에 있어서도 KoBART 가 한국어 40G 의 데이터로 학습되어 있어 성능 차이가 발생한것으로 판단된다.

5. 결론 및 향후 연구과제

최근 우수한 성능을 보이고 있는 사전학습 기반의 BART 모델이 도메인에 상관없이 우수한 성능을 보이는 것으로 실험결과로 알 수 있었다. 특히 학습 데이터셋은 원본 문장에 대한 요약이 생성기법으로 만들어져 있어 대용량의 Corpus 로 사전학습된 BART 기반의 생성형 모델이 TextRank 나 BERT 기반의 추출모델에 비해 성능이 우수한것으로 파악되었다.

추후 진행할 연구과제로 사전학습된 기본 모델에 Prompt tuning[12]을 사용하는 사후학습레이어(Post Trained Layer)를 적용하여 성능평가를 시도해 볼 예정이며 법률 문장의 특징을 분석하여 전처리과정을 추가하여 진행할 예정이다.

특히, 법률 문장내 범조항등의 주요한 키워드에 대해 우수한 성능의 사전학습 모델이라 하더라도 누락되는 경우가 있으며 요약결과를 전문가가 확인해야하는 번거로움이 있는바 법률 문서 구조기반의 특화된 모델개발을 향후 연구과제로 진행예정이다.

참고문헌

[1] Zhong, Haoxi, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. "How does NLP benefit legal system: A summary of legal artificial intelligence." arXiv preprint arXiv:2004.12158 (2020).
 [2] 배주호, 박석."LEXAI : 설명 가능한 인공지능을 이용한 법률 문서 유사도 분석 서비스."정보과학회 논문지 47.11(2020):1061-1070.
 [3] 김용환. "법률정보시스템을 위한 텍스트 마이닝 적용 방안 - 명예 훼손 판례를 대상으로 -" 한국문헌정보학회지 54, no.1 (2020) : 387-409.

[4] Sizov, G. "Extraction-Based Automatic Summarization: Theoretical and Empirical Investigation of Summarization Techniques." (2010).
 [5] Mihalcea, Rada, and Paul Tarau. "TextRank: Bringing order into text." In Proceedings of the 2004 conference on empirical methods in natural language processing, pp. 404-411. 2004.
 [6] Deepali Jain, Malaya Dutta Borah, and Anupam Biswas. 2020. Fine-Tuning TextRank for Legal Document Summarization: A Bayesian Optimization Based Approach. In Forum for Information Retrieval Evaluation (FIRE 2020)
 [7] Nallapati, Ramesh, Bowen Zhou, Caglar Gulcehre, and Bing Xiang. "Abstractive text summarization using sequence-to-sequence rnns and beyond." arXiv preprint arXiv:1602.06023 (2016).
 [8] Gu, Jiatao, Zhengdong Lu, Hang Li, and Victor OK Li. "Incorporating copying mechanism in sequence-to-sequence learning." arXiv preprint arXiv:1603.06393 (2016).
 [9] Liu, Yang. "Fine-tune BERT for extractive summarization." arXiv preprint arXiv:1903.10318 (2019).
 [10] Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. "Bart: Denoising sequence to-sequence pre-training for natural language generation, translation, and comprehension." arXiv preprint arXiv:1910.13461 (2019).
 [11] Chin Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out Workshop held in conjunction with ACL'2004, pages 74–81, Barcelona, Spain, 25–26 July 2004.
 [12] Liu, Xiao, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang and Jie Tang. "GPT Understands, Too." ArXiv abs/2103.10385 (2021):