공학석사학위논문

어텐션 메커니즘을 활용한 특허 문서의 다중 레이블 분류

Multi-label Patent Classification with Attention Mechanism

2019 년 2 월

서울대학교 대학원

산업공학과

박 노 일

어텐션 메커니즘을 활용한 특허 문서의 다중 레이블 분류

Multi-label Patent Classification with Attention Mechanism

지도교수 조 성 준

이 논문을 공학석사 학위논문으로 제출함

2019 년 2 월

서울대학교 대학원

산업공학과

박 노 일

박노일의 공학석사 학위논문을 인준함

2018 년 12 월

위 원 장 박 종 헌 (인)

부위원장 조 성 준 (인)

위 원 장 우 진 (인)

초록

전 세계적으로 지적 재산권에 관한 특허 출원은 계속해서 증가하는 추세이다. 하지만 특허 심사는 여전히 소수의 전문적인 지식을 갖춘 심사관들에 의존하고 있기 때문에 특허청의 등록 승인을 받는데까지 긴 시간이 걸리고 있다. 따라서 방대한 양의 특허 정보를 기술적 분야에 따라 자동적으로 분류하는 방법에 대한 연구가 활발히 이루어져 왔다. 본 연구에서는 최근 컴퓨터 비전에 이어 자연어 처리에서도 널리 사용되고 있는 딥러닝 알고리즘을 통해 특허 문서의 다중 레이블 분류 문제에 접근하고자 한다. 구체적으로 GRU 기반의 문서 인코더와 어텐션 메커니즘을 활용하여 특허 문서의 국제특허분류(IPC) 코드를 예측하는 모델을 제안한다. 제안하는 모델의 학습과 평가를 위해 앞선 연구에서 사용한 특허 문서 데이터셋 USPTO-2M을 사용한다. 정밀도(Precision), 재현율(Recall), F_1, F_β 점수를 통해 평가한다. 또한 어텐션 메커니즘을 통해 특허 문서의 분류 결과에 대한 단어별 영향력을 분석하여 키워드를 탐색한다. 특히 특허 문서의 단어별 어텐션 스코어의 시각화를 통해 분류 결과에 대한 기여도를 단어 단위로 비교하고 비중이 높은 단어를 키워드로 선별할 수 있다. 이를 통해 향후 특허 분석이나 키워드 검색에서 활용할 수 있는 의의를 갖는다.

주요어: 특허 분류, 다중 레이블 분류, 어텐션 메커니즘, Gated Recurrent Unit, 문서 인코더, 단어 임베딩

학번: 2017-22964

목차

표 목차

그림 목차

제 1 장 서론

최근 수년 간 세계 지적 재산권(IP) 출원은 꾸준하게 증가하고 있는 추세이다. 세계지식재산기구(WIPO)가 발표한 2017 세계 지적 재산지표 보고서[47]에 따르면 전 세계적으로 특허 출원 건수는 2015년 2,887,300건에서 2016년 3,127,900건으로 8.3% 증가하였다. 상표권 출원 활동 또한 2016년을 기준으로 1년 사이에 13.5% 증가하였고 이와 같은 상승세를 7년 연속으로 이어가고 있다. 하지만 이처럼 빠르게 늘어나는 특허 출원 요청에 비해, 이를 심사하고 정식으로 등록하는데 소요되는 시간은 여전히 더딘 모습을 보여주고 있다. 특허 출원일로부터 특허청 심사관의 심사 결과를 통지 받고 특허 등록 여부가 결정되는 데까지 소요되는 특허 출원 기간은 2016년도를 기준으로 대한민국 특허청은 약 16.2개월, 미국특허청은 약 22.6개월, 유럽특허청은 약 23.3개월이 소요되고 있다. 또한 특허 심사 및 등록을 앞두고 계류 중인 특허의 양도 꾸준하게 증가하고 있다.

이처럼 전 세계적으로 특허 출원 및 등록에 대한 수요는 계속 증가하고 있지만, 그에 비해 출원된 특허를 심사하고 처리하는 속도는 이를 따라가지 못하고 있다. 이는 출원된 특허를 IPC (International Patent Classification)와 같은 특허분류체계에 따라 나누고 적합성을 심사하는 일이 전문적인 지식을 갖춘 소수의 특허 심사관들에 의해 수작업으로 이루어지고 있기 때문이다. 따라서 특허 심사에 소요되는 시간과 비용을 줄이고 특허 출원을 보다 효율적으로 처리하기 위해서 특허 문서의 분류 자동화가 필수적이다. 본 논문에서는 대량의 특허 문서에 대한 국제특허분류(IPC) 코드를 예측하여 분류하는 문제를 다루고자 한다. 1.1 절에서 이 문제에 대해 자세히 소개한다.

1.1 연구 배경 및 동기

특허 문서의 분류 문제는 기존의 문서 분류 중에서 도메인 특화된 연구 분야로써 다양한 머신러닝 기법들을 통한 접근이 이루어져왔다. 특히 k-Nearest Neighbor (k-NN)[14], Support Vector Machine (SVM)[14, 48, 13], Naive Bayes[14, 13], k-means clustering[21], TF-IDF[31] 등의 다양한 분류 알고리즘을 통해 특허 문서를 주제에 따라 군집화하고 분류하는 방법이 제시되었다. 이는 기존의 자연어처리 분야에서 널리 사용되는 문서 분류 모델을 특허 데이터에도 적용하는 형태로 볼 수 있다. 최근에는 Convolution Neural Network (CNN), Recurrent Neural Network (RNN) 등의 딥러닝 알고리즘을 컴퓨터 비전뿐만 아니라 자연어처리 분야에서도 활용하여 감성 분류[26, 20, 19, 43, 22], 기계 번역[5], 질의 응답[42, 22, 17], 개체명 인식[12, 9, 24] 등의 과제에서 좋은 성능을 보여준 연구들이 있었다. 기존의 머신러닝 기법은 수작업으로 직접 추출한 피처에 강하게 의존하여 많은 시간과 비용이 소요되는 것이 불가피하였다. 하지만 최근 수 년간 제안된 딥러닝 알고리즘은 피처 추출에서부터 모델의 학습과 추론까지의 모든 과정을 자동화하였고, 효율적인 알고리즘을 통해 크기가 큰 데이터를 빠르게 학습할 수 있게 하였다[53]. 따라서 본 연구에서도 대량의 특허 문서를 효과적으로 분류하기 위하여 딥러닝 알고리즘을 적용하고자 한다.

1.2 연구 목적 및 문제 정의

본 연구에서는 Gated Recurrent Unit (GRU)[10] 기반의 문서 인코더와 어텐션 메커니즘[5]을 활용하여 대량의 특허 문서를 자동으로 분류하고자 한다. 특허 문서의 특성상 하나의 문서가 동시에 다수의 레이블에 속할 수 있다. 따라서 본 연구는 특허 문서의 다중 레이블 분류에 관한 방법을 제안하고자 한다. 이를 위해 특허 문서의 제목, 초록 항목의 텍스트 정보와 함께 IPC 코드를 문서의 레이블 정보로 활용한다. 즉

텍스트 정보는 단어 벡터 시퀀스 X로 임베딩하고 레이블 정보는 Multi-hot 벡터 Y로 인코딩하여 특허 문서 분류 모델 $f(Y|X)$를 학습시킨다. 이를 통해 새로운 특허 문서 $\hat{X}$가 주어졌을 때, 특허 문서 분류 모델을 통해 해당 문서가 IPC 클래스 Y에 속할 확률 $f(Y|\hat{X})$을 예측하는 것을 목표로 한다.

본 연구의 공헌은 크게 세 가지로 정리할 수 있다. 첫 번째, 제안하는 특허 분류 모델을 통해 대량의 특허 정보를 효과적으로 분류할 수 있기 때문에 향후 특허 심사에 소요되는 시간과 비용을 획기적으로 줄일 수 있다. 특히 불용어(Stopwords) 제거와 원형 복원(Lemmatization)과 같은 최소한의 전처리만으로도 약 2백만 건의 특허 문서에 대한 피처 벡터를 추출할 수 있기 때문에 효율적인 분류 모델이라고 판단된다. 두 번째, 어텐션 메커니즘[5]을 통해 분류 결과에 큰 영향을 미치는 키워드를 선별하여 보여줄 수 있다. 이를 통해 분류 결과에 대한 논리적인 해석을 제공하여 향후 특허 심사 또는 분석에서 보조하는 역할로써 활용할 수 있을 것으로 기대된다. 세 번째, IPC의 서브클래스 레벨의 분류뿐만 아니라 목적에 따라 다양한 종류와 레벨의 분류체계에 대한 특허 문서 분류 모델을 확장할 수 있다.

1.3 논문 구성

본 논문은 5 장으로 구성된다. 제 2장에서는 특허 문서 분류와 관련된 선행 연구를 살펴본다. 제 3장에서는 본 연구가 제안하는 특허 문서의 다중 레이블 분류 모델을 소개한다. 제 4장에서는 실험에서 사용한 데이터를 소개하고 실험의 결과를 분석한다. 마지막으로 제 5장에서는 본 연구의 결론과 향후 연구 방향을 제시한다.

제 2 장 관련 연구

2.1 특허 문서의 특징

특허 문서는 발명자에게 부여되는 독점적 권리인 지적재산권을 정의하기 위한 목적을 가진다. 일반적인 과학기술 분야의 논문과 다른 특허 문서만의 교유한 특징이 있기 때문에, 정확한 특허 문서 분류를 위해 반드시 고려해야한다. 특허 문서 내 항목들의 구성은 국가별로 차이가 있지만 본 연구에서 다루는 미국특허를 기준으로 소개한다. 특허 문서는 크게 서지정보, 제목(Title), 초록(Abstract), 배경(Background), 요약(Summary), 도면(Drawings), 명세서(Descriptions), 청구항(Claims) 등의 항목으로 구성되어 있다. 이 중 서지정보는 발명권자(Inventors), 특허권자(Applicant), 패밀리 아이디, 출원번호(Application Number), 국제특허분류(IPC) 코드, 국내외 참조특허 등의 세부 항목들로 구성되어 있다. 데이터의 형태를 기준으로 나누면 제목, 배경, 요약, 명세서, 청구항과 같은 텍스트 데이터와 도면과 같은 이미지 데이터로 구성된다. 본 연구에서는 이 항목들 중에서 제목과 초록 항목의 텍스트 데이터와 함께 IPC 코드를 문서의 레이블 데이터로 사용한다.

IPC는 국가마다 서로 다른 특허분류체계를 국제적으로 통일하기 위해 1968년에 도입된 국제 특허 분류체계이다. WIPO에서 발표한 “Guide To The International Patent Classification (2018)”[4]에 따르면 IPC는 총 8개의 섹션, 130개의 클래스, 640개의 서브클래스, 7,400개의 메인그룹, 그리고 약 72,000개의 서브그룹으로 나눠지는 계층적인 구조로 이루어져 있다. 최상위 레벨인 섹션은 표2.1와 같이 특허 문서가 포함하는 기술적 주제에 따라 알파벳 A부터 H까지 표기된 코드로 분류된다. 두 번째 레벨인 클래스는

해당 섹션의 세부적인 주제에 따라 두 자리의 숫자로 표기된 코드로 분류된다. IPC의 계층적 구조에 관한 예시를 들면 다음과 같다. “G06T 1/40” IPC 서브그룹 레이블은 신경망(Neural Networks) 기술 관련 특허 문서에 대한 범주로써, 순서대로 “물리학”에 관한 섹션 G, “컴퓨팅; 계산; 산출”에 관한 클래스 G06, “이미지 데이터 처리 또는 생성”에 관한 서브클래스 G06T, “범용적 이미지 데이터 처리”에 관한 그룹 G06T 1에 속한다. 특허 문서를 기술적 주제에 따라 체계적으로 정리하고 관리하기 위한 기준이기 때문에 본 연구에서는 IPC를 특허 문서의 레이블 정보로 활용하기에 적절하다고 가정하였다.

표 2.1: IPC 코드 섹션별 분류

IPC 코드분류	내용
A 섹션	생활필수품
B 섹션	처리조작; 운수
C 섹션	화학; 야금
D 섹션	섬유, 지류
E 섹션	고정구조물
F 섹션	기계공학; 조명; 가열; 무기; 폭파
G 섹션	물리학
H 섹션	전기

표 2.2: IPC Scheme 예시

섹션	G	물리학
클래스	G06	컴퓨팅; 계산; 산출
서브클래스	G06C	모든 계산이 기계적으로 행하여지는 디지털 컴퓨터
	G06D	디지털 유체압 계산장치
	G06E	광학 계산 장치
	G06F	전기에 의한 디지털 데이터처리
	G06G	아날로그 컴퓨터
	G06J	하이브리드 계산장치
	G06K	데이터의 인식; 데이터의 표현; 기억 장치; 기억장치
	G06M	계수메커니즘; 다른 방식으로는 분류되지 않는 계수
	G06N	특정 계산 모델에 기반한 컴퓨터 시스템
	G06Q	관리용, 상업용, 금융용, 경영용, 감독용 또는 예측용으로 특히 적합한 데이터 처리 시스템 또는 방법; 그 밖에 분류되지 않는 관리용, 상업용, 금융용, 경영용, 감독용 또는 예측용으로 특히 적합한 시스템 또는 방법
	G06T	이미지 데이터 처리 또는 생성

2.2 단어 임베딩

자연어 처리 분야에서는 컴퓨터로 인간의 언어를 이해하고 분석하는 기법들을 연구해왔다. 특히 비슷한 분포를 가진 단어들은 비슷한 의미를 가진다는 Distributional Hypothesis를 바탕으로 단어의 수치적 표현 방법에 대한 연구가 계속 되었다. 단어가 가진 의미를 연속적인 실수 값을 가진 벡터로 표현하는 아이디어는 1980년대 중반부터 시작되었다[40]. 이러한 개념은 2000년대에 들어 인공 신경망을 통해 단어 임베딩을 학습하는 언어 모델링 구조에 대한 연구로 이어졌다[6, 11, 12]. Bengio et al.은 단어에 대한 분산 표현 학습을 통해 문맥에서 단어 표현의 고차원성을 줄이기 위한 확률론적 언어 모델을 제시하였다[6]. Mikolov et al.[36]은 기존의 언어 모델을 RNN 구조로 확장하여 적용시켰다. 2013년에 등장한 Word2vec[34] 방법론은 기존의 인공신경망 기반 학습방법과 달리, 은닉층에서의 비선형성을 제거하고 단어 간 projection 층을 공유하여 계산복잡도를 획기적으로 줄였다. Word2vec[34] 모델에는 학습을 위한 네트워크

모델 두 가지, ‘CBOW’와 ‘Skip-gram’이 있는데, 전반적인 성능에 있어서 Skip-gram이 우세한 것으로 나타났고 이후의 연구에서도 비슷한 양상을 보였다[37][52].

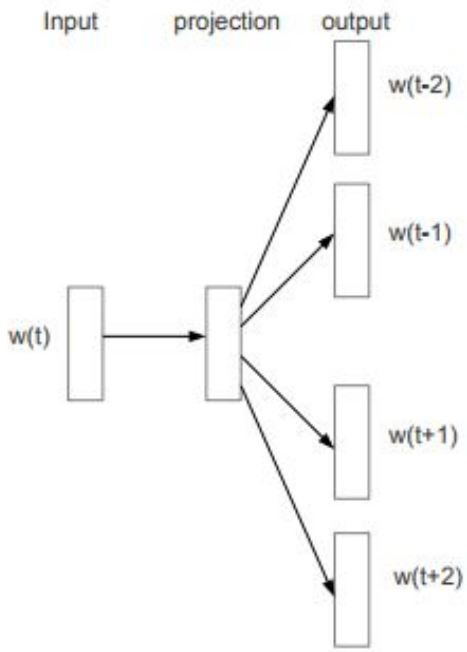

그림 2.1: Skip-gram method[34]

Skip-gram 모델[34]은 그림 2.1과 같이 인공신경망 구조를 통해 주어진 단어 w_t 주변의 문맥을 파악하여 단어 임베딩 벡터를 생성한다. 여기서 w_t 주변에 등장할 수 있는 단어들을 예측하는데, 주어진 단어와 가까울수록 관련성이 더 높을 것이라고 가정하여 다음과 같이 확률을 계산한다.

$$p(w_O|w_I) = \frac{\exp({v'_{w_O}}^T v_{w_I})}{\sum_{w=1}^{W} \exp({v'_w}^T v_{w_I})} \tag{2.1}$$

그리고 현재 단어와의 거리가 멀수록 낮은 확률로 택하는 Negative Sampling 방법을 통해 모델의 계산복잡도를 낮추었다. 이를 통해 Skip-gram 모델[34]은 T개의 연속된

단어 시퀀스 $w_1, w_2, w_3, ..., w_T$에 대해서 다음과 같은 로그 확률을 최대화하는 것을 목표로 학습한다.

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq l\leq c, j\neq 0}\log p(w_{t+j}|w_t) \tag{2.2}$$

식 2.2에서 c는 학습 문맥(context)의 크기를 의미하고, 그 크기가 커질수록 학습 데이터가 방대해지기 때문에 시간은 오래걸리지만 그만큼 정확도를 높여준다[34].

2.3 문서 분류

Kim[20]은 2.2 절에서 소개한 Word2vec[34]을 통해 문서 벡터의 초기값을 만들고 Multi-channel CNN 모델을 통해 문서 분류 실험에서 좋은 성능을 보여주었다. 이는 컴퓨터 비전에서 주로 사용되던 CNN 분류 모델[27]이 자연어처리 분야에서도 널리 사용되는 계기가 되었다. Kalchbrenner et al.[19]은 k-max 풀링 레이어를 적용한 Deep CNN 모델을 통해 문서 분류 문제에 접근하였다. Johnson et al.[18]은 고차원의 one-hot 벡터를 입력값으로 사용하는 방법을 제안하였고 이는 문서 분류 모델에 있어 괜찮은 성능을 보여주었다. 이처럼 단어 단위의 문서 분류 모델 이외에도, 글자 단위에 CNN 모델을 적용한 연구가 있었다. Zhang et al.[54]은 Character-level CNN 모델을 제안하였고 이는 문서 분류 과제에서 좋은 성능을 보여주었다.

Socher et al.[41]과 Tai et al.[43]은 문장의 구조적 특징에서 착안한 트리 구조의 Recursive Networks 모델을 활용하였다. Socher et al.[41]은 Recursive Neural Tensor Networks를 통한 문서 분류 방법론을 제안하였다. Tai et al.[43]은 문장의 구조를 바탕으로 tree-structured LSTMs 모델을 제안하였다. 이밖에도 LSTM과 CNN 구조를 함께 사용하여 문장 또는 문서 분류 모델을 제안한 연구들[23, 55]이 있었다. Tang et al.[44]은 계층적 구조를 통해 문서의 감정 분류 과제에 접근하였다. CNN이나 LSTM을 통해 문장 벡터를 추출하였고 이를 bi-directional gated RNN 모델을 통해 문서 벡터로

변환하였다. 문서 분류 과제뿐만 아니라 문장 생성[29]과 언어 모델링[32]에서도 문서의 계층적 구조를 모델에 활용한 사례가 있다.

어텐션 메커니즘은 Bahdanau et al.[5]의 기계번역 연구에서 제시되었다. 인코더 디코더 구조와 함께 사용된 어텐션 메커니즘[5]은 기준이 되는 언어의 단어와 이에 대응되는 번역 후의 단어들을 효과적으로 선별하였다. Xu et al.[49]은 이미지 캡션 생성 방법론에 어텐션 메커니즘을 적용하였는데, 캡션의 단어를 생성할 때 이와 관련이 되는 이미지의 부분을 선택적으로 보여줄 수 있었다. Yang et al.[51]은 문서 분류를 위한 Hierarchical Attention Network를 제안하였다. GRU 기반의 인코더에 어텐션 메커니즘을 결합하여 단어 벡터로부터 문장 벡터와 문서 벡터를 단계적으로 추출하였고, 이를 통해 문서 분류에서 좋은 성능을 보여주었다. 이밖에도 구문 분석[45], 자연어 질의응답[42][22][17], 이미지 질의응답[50] 등의 연구에서도 어텐션 메커니즘이 활용되었다. 본 연구는 Hierarchical Attention Network[51]의 구조에 착안하여 특허 문서의 다중 레이블 분류 방법론을 제안하고자 한다.

2.4 특허 분류

특허 문서 분류에 관한 연구는 일반적인 문서 분류에서 널리 사용되는 SVM, k-NN, 인공신경망 등의 다양한 머신러닝 알고리즘을 특허 데이터에 알맞게 응용하는 방식으로 이루어져왔다[7]. Larkey[25]는 k-NN을 사용하여 미국특허분류 코드 USPC에 대한 특허 문서 분류 연구를 하였다. 제목, 초록, 배경, 요약 항목에서 처음 20개의 문장과 청구항 항목 전체를 사용하였을 때 분류 모델의 성능이 가장 좋다는 것을 보여주었다. Fall et al.[14]은 Naive Bayes, k-NN, SVM, SNoW와 같은 다양한 분류 모델을 사용하여 특허 문서의 분류에 대한 연구를 하였다. 서브클래스 레벨에서의 IPC 분류를 위해 특허 문서의 제목, 초록, 청구항 항목에 대하여 처음부터 300번째 단어까지를 사용하였다.

연구에 따르면 Top-prediction과 All-category 평가에서 SVM이 가장 좋은 성능을 보였고, Three-guesses 평가에서는 k-NN이 가장 좋은 성능을 보여주었다. Wu et al.[48]은 반도체 장비에 관한 특허 문서 234건으로부터 키워드를 추출하여 유전 알고리즘 기반의 SVM 분류 모델(HGA-SVM)을 제안하였다. 이는 기본적인 SVM 모델에 비해 향상된 82%의 분류 정확도를 보여주었다. Guyot et al.[16]은 2,000건의 특허 문서에 대해 Winnow 알고리즘 기반의 분류 모델을 제안하였다. Li et al.[31]은 2층 순전파 인공신경망과 Levenberg-Marquardt 알고리즘을 활용한 분류 모델을 제안하였다. 미국특허분류코드 USPC 360/324에 해당되는 1,948개의 특허 문서에 대해 학습하였고 77.12%의 정확도를 보여주었다. 하지만 Wu et al.[48]과 Guyot et al.[16], Li et al.[31]의 연구는 적은 양의 데이터셋을 사용하였기 때문에 대량의 특허 문서에 적용할 수 없다는 한계가 있다. Chen et al.[8]은 SVM, k-means clustering, k-NN 알고리즘과 TF-IDF를 통해 3단계 분류 모델(TPC)을 제안하였다. 21,104개의 특허 데이터에 대해 서브그룹 레벨에서 레이블을 예측하였을 때 36.07%의 정확도를 보여주었다. 최근에는 Li et al.[30]이 단어 임베딩과 CNN 구조를 활용하여 IPC 서브클래스 레벨에 대한 특허 분류 모델을 제안하였다. CLEF-IP[38]와 USPTO-2M[30] 데이터셋을 사용하여 학습한 모델은 각각 83.50%, 73.88%의 정밀도를 기록하였다.

제 3 장 제안하는 방법

본 연구는 Yang et al.의 연구[51]를 바탕으로 특허 문서의 다중 레이블 분류 모델을 제안하고자 한다. 구체적으로 GRU 기반의 문서 인코더와 어텐션 메커니즘[5]을 통해 특허 문서의 국제특허분류(IPC) 코드를 예측하는 것을 목표로 한다. 특허 문서 분류 모델의 전체적인 프레임워크는 그림 3.1과 같다.

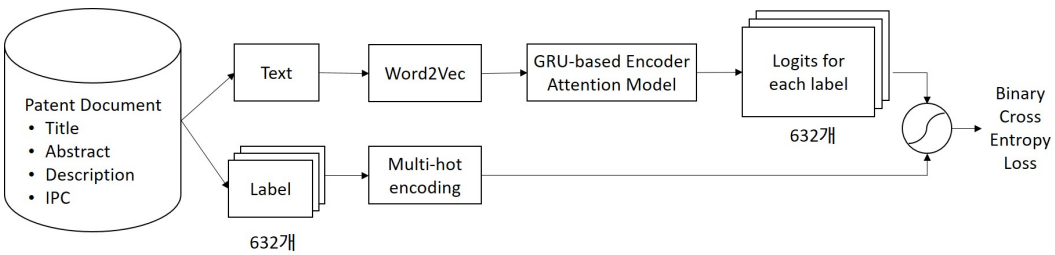

그림 3.1: 제안하는 방법의 전체적인 프레임워크

먼저 특허 문서의 제목, 초록, 명세서 부분의 텍스트 정보와 IPC 코드를 추출한다. 텍스트 정보는 단어 시퀀스로 이루어져 있으며 Word2vec[34]을 통해 단어 벡터 시퀀스로 임베딩한다. IPC 코드는 다중 레이블 분류를 위해 레이블의 총 개수를 길이로 갖는 Multi-hot 벡터로 인코딩한다. 임베딩된 단어 벡터 시퀀스와 레이블 벡터를 입력값으로 하는 GRU 기반의 문서 분류 모델을 학습한다. Binary Cross Entropy 함수를 사용하여 각각의 레이블에 대해 비교하여 손실 함수를 구한다. 제안하는 문서 분류 모델은 이러한 손실 함수의 값을 줄여나가는 방향으로 학습한다. 최종적으로는 새로운 특허 문서가 주어졌을 때, 텍스트 정보만을 활용하여 서브클래스 레벨에 해당하는 IPC 코드를 예측하는 것을 목표로 한다.

3.1 어텐션 기반 특허 문서 분류 모델

3.1.1 GRU 기반 단어 시퀀스 인코더

GRU[10]는 Recursive Neural Network (RNN) 계열의 대표적인 셀 가운데 하나이다. RNN의 고질적인 문제점 중 하나인 Gradient Vanishing and Explosion 문제를 극복하기 위해 크게 두 가지 게이트, Update gate와 Reset gate를 사용한다. GRU는 기존 Long Short-Term Memory (LSTM) 셀의 장점을 유지하면서도 Forget gate를 사용하지 않음으로써 계산복잡도를 낮추었다.

GRU를 구성하는 Reset gate r_t 와 Update gate z_t 는 각 시점의 정보를 어떤 식으로 업데이트할지 결정하는 역할을 한다. 현 시점 t를 기준으로 다음 상태를 업데이트하는 식은 다음과 같이 정의된다.

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{3.1}$$

식 3.1은 이전 시점의 상태 h_{t-1}와 현 시점의 새로운 상태 $\tilde{h}_t$를 선형보간한 결과이다. 여기서 $\odot$는 성분별 곱셈을 뜻한다. Update gate z_t는 얼마나 과거의 정보를 보존하고 새로운 정보를 더할지 결정하는 역할을 한다. z_t는 다음과 같이 업데이트 된다.

$$z_t = \sigma \left(W_z x_t + U_z h_{t-1} + b_z \right) \tag{3.2}$$

여기서 W_z와 U_z는 각각 현 시점의 입력값인 x_t와 이전 시점의 은닉층 값인 h_{t-1}를 선형결합하는 매개변수이다. 현 시점의 후보 상태 $\tilde{h}_t$는 기존 RNN[10]과 유사한 방식으로 업데이트 된다.

$$\tilde{h}_t = \tanh \left(W_h x_t + r_t \odot (U_h h_{t-1}) + b_h \right) \tag{3.3}$$

식 3.3에서 r_t는 Reset gate로써, 과거의 상태가 후보 상태에 얼마나 영향을 미치게 할지 결정하는 역할을 한다. 활성화 함수로는 하이퍼볼릭 탄젠트 함수가 사용되므로 $\tilde{h}_t$는 -1 부터 1 사이의 값을 갖는다. 여기서 현 시점 t의 정보는 r_t 값에 상관없이 반영된다. Reset gate r_t는 다음과 같이 계산된다.

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1} + b_r\right) \tag{3.4}$$

식 3.4의 활성화 함수는 시그모이드 함수이므로 r_t는 0 부터 1 사이의 값을 갖는다. r_t가 0일 때는 이전 시점의 정보를 모두 잊게 되고, 1일 때는 이전 시점의 정보를 모두 기억하게 된다.

3.1.2 어텐션 기반 특허 문서 인코더

하나의 특허 문서를 T개의 연속된 단어 $w_i, i \in [1, T]$로 이루어진 단어 시퀀스라고 할 때, 특허 문서 분류 모델은 그림 3.2와 같은 구조로 정의한다. 앞서 3.1.1 절에서 정의한 GRU 기반 시퀀스 인코더의 입력값을 위해 단어 시퀀스를 Word2vec[34]을 통해 벡터 공간에 임베딩한다.

$$x_t = W_e w_t, t \in [1, T]$$

$$\overrightarrow{h}_t = \overrightarrow{GRU}(x_t), t \in [1, T]$$

$$\overleftarrow{h}_t = \overleftarrow{GRU}(x_t), t \in [T, 1]$$

$$h_t = \left[\overrightarrow{h}_t, \overleftarrow{h}_t\right], t \in [1, T]$$

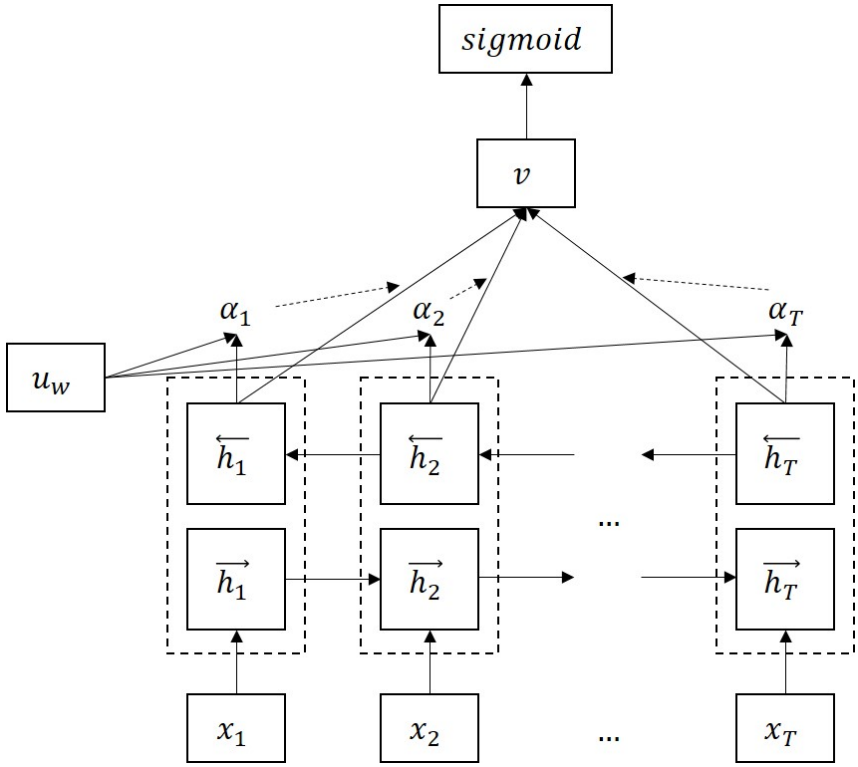

그림 3.2: 특허 문서 분류 모델의 구조

여기서 W_e는 단어 임베딩의 가중치 벡터이다. 임베딩된 단어 벡터 시퀀스 x_t를 GRU 인코더의 입력값으로 넣는다. 이 때, 단어 벡터 x_t와 앞, 뒤 단어들 간의 문맥을 반영하기 위해 양방향(bi-directional) GRU 인코더를 사용한다. h_t는 양방향 GRU 인코더를 통해 얻은 두 개의 은닉 벡터를 연결(concatenate)한 벡터로써 문서 내에서 t 번째 단어에 관한 정보를 요약하고 있다.

$$u_t = \tanh\left(W_w h_t + b_w\right) \tag{3.5}$$

$$\alpha_t = \frac{\exp\left(u_t^T u_w\right)}{\sum_t \exp\left(u_t^T u_w\right)} \tag{3.6}$$

$$v = \sum_{t} \alpha_t h_t \tag{3.7}$$

다음은 어텐션 메커니즘[5]을 활용하여 벡터 $h_t, t \in [1, T]$로부터 문서 벡터 v를 인코딩하는 과정이다. 먼저 식 3.5와 같이 하이퍼볼릭 탄젠트 함수를 활성화 함수로 하는 한 층짜리 MLP를 통해 h_t를 벡터 u_t로 변환한다. 소프트맥스 함수를 통해 u_t와 단어 문맥 벡터 u_w의 유사도를 정규화하여 어텐션 스코어 α_t를 얻는다. 여기서 $\alpha_t \in [0, 1]$는 각 단어가 해당 특허 문서의 주제를 결정하는데 얼마나 영향을 주는지 반영한 가중치 값을 의미한다. 식 3.6의 u_w는 단어의 문맥 벡터로써, 모델 학습시 무작위 값으로 초기화하여 함께 업데이트하는 매개변수이다. 최종적으로 식 3.7로부터 각 단어별 은닉 벡터 h_t에 어텐션 스코어 α_t를 곱한 단어별 가중치 합인 v를 문장 벡터로 얻을 수 있다. 즉 v는 문서의 전체적인 정보를 담고 있는 피처 벡터로써 양방향 GRU 인코더의 은닉층 차원 만큼의 길이를 가진다.

3.1.3 특허 문서 분류

앞서 3.1.2 절에서 얻은 문서 벡터 v는 특허 문서의 모든 단어들로부터 얻은 상위 레벨의 피처 벡터이다. 이 때, 다중 레이블 분류의 특성상 특허 문서가 각 레이블에 속할 확률이 서로 독립이므로 다음과 같이 확률을 계산한다.

$$p_j = \sigma\left(W_c v + b_c\right), j \in [1, N] \tag{3.8}$$

식 3.8은 모든 레이블에 대한 확률을 독립적으로 예측하기 위하여 시그모이드 함수를 활성화 함수로 사용한다. p_j는 j번째 레이블에 대한 특허 문서의 확률 예측값으로써, 0.5를 기준으로 j번째 레이블에 대한 이항 분류를 한다. 즉, p_j가 0 이상 0.5 미만의 값일 때는 j번째 레이블에 속하지 않고, 0.5 이상 1 이하의 값일 때는 j번째 레이블에

속한다고 판단한다.

$$\mathcal{L} = -\left[\beta t_j \log p_j + (1 - t_j) \log(1 - p_j)\right] \tag{3.9}$$

손실 함수 $\mathcal{L}$은 Binary Cross Entropy 함수를 사용한다. 식 3.9에서 t_j는 앞서 multi-hot 인코딩한 레이블 벡터의 j번째 레이블에 대한 이진값을 나타낸다. β는 positive weight로써 IPC의 서브클래스 레벨에서의 다중 레이블 분류에 있어, 참 긍정(true positive)에 비해 참 부정(true negative)의 개수가 지나치게 많은 상황을 보정해주기 위한 하이퍼파라미터이다. β의 값이 클수록 정밀도가 감소하고 재현율이 증가하며, 이는 통계학적으로 1종 오류에 비해 2종 오류에 대한 가중치를 크게 주는 효과가 있다. 즉, 1보다 큰 β 값을 통해 특허 문서 분류 모델의 예측값을 참(Positive)값이 되도록 유도하여 학습한다. β에 따른 실험 구성과 평가는 4.2 절에서 후술하였다. 모델에서 사용한 변수와 매개변수에 대한 정의는 다음 표 3.1로 정리하였다.

표 3.1: 모델의 변수 정의

변수	정의
w_t	t번째 단어에 대한 one-hot 벡터, $t \in [1, T]$
W_e	단어 임베딩 가중치 행렬, W_e
x_t	t번째 단어의 임베딩 벡터, $t \in [1, T]$
h_t	t번째 단어에 대한 GRU 인코더의 은닉층 벡터, $t \in [1, T]$
W_w, b_w	h_t에 대한 1층 MLP의 매개변수 행렬
u_t	h_t에 대한 1층 MLP의 은닉층 벡터
u_w	단어 시퀀스의 문맥 벡터
α_t	t번째 단어의 어텐션 스코어(스칼라), $t \in [1, T]$
v	문서 벡터
β	Positive weight(스칼라)

제 4 장 실험 결과 및 분석

4.1 데이터셋

본 연구는 특허 분류 과제를 위한 벤치마크 데이터셋 USPTO-2M[30]을 실험 데이터로 사용하였다. USPTO-2M[30] 데이터셋은 2006년부터 2015년까지의 미국특허청에 등록된 특허 2,000,147건으로 구성되어있다. 각 특허 문서는 제목, 초록, 출원번호, 서브클래스 레벨의 IPC 레이블에 관한 정보가 포함되어 있다. 그림 4.1과 같이 USPTO-2M[30] 데이터셋의 제목과 초록 항목은 특수문자와 숫자가 제거된 영어 알파벳 소문자로 구성된 텍스트 데이터이다. 제안하는 분류 모델의 학습과 검증, 시험을 위해 표 4.1와 같이 데이터를 나누었다. 편의상 특허 문서를 연도별로 나누어 2006년부터 2013년까지의 특허 문서를 학습 데이터로, 2014년도의 특허 문서를 검증 데이터로, 2015년도의 특허 문서를 시험 데이터로 구성하였다.


```
{
    "Subclass_labels": [
      "B64D",
      "B64G"
    ],
    "Abstract": "a method of countering the effects of g forces on a person comprises passing
inflation gas to a bladder around a leg of the person by starting the inflation of the bladder to
an operating pressure from a point adjacent an ankle of the leg as well as from a point spaced
from said ankle towards the abdomen of the wearer the inflation then progresses away from
the ankle at the same time as the upper part of the bladder is being inflated when the bladder
is inflated to the operating pressure the blood vessels of the leg are constricted",
    "Title": "aircrew ensembles",
    "No": "US08925112"
}
```


그림 4.1: USPTO-2M[30] 데이터셋 예시

표 4.1: USPTO-2M[30] 데이터셋 특성

	학습 데이터	검증 데이터	시험 데이터
특허 문서 개수	1,645,175	303,332	49,900
서브클래스 레벨의 IPC 코드 개수	632	622	606
특허 문서당 IPC 코드 평균 개수	1.28	1.54	1.93
특허 문서당 IPC 코드 최대 개수	18	16	13
특허 문서당 IPC 코드 최소 개수	1	1	1

4.2 평가 방법

특허 문서의 다중 레이블 분류 모델의 성능을 평가하기 위해, 다음과 같이 정밀도(Precision), 재현율(Recall), F_1, F_β를 사용하였다. 특허 문서 한 개가 다수의 레이블에 속할 수 있는 문제의 특성에 따라, 본 연구에서 제안하는 특허 문서 분류 모델은 주어진 특허 문서가 속할 수 있는 레이블에 대해 독립적으로 확률을 예측한다. 주어진 특허 문서에 대한 예측 레이블과 실제 레이블을 각각 그림 4.2의 Prediction과 True Label이라고 할 때, 정밀도는 참이라고 예측한 레이블 중에서 실제로 참값을 갖는 레이블의 비율이다. 재현율은 실제로 참값을 갖는 레이블 중에서 참이라고 예측한 레이블의 비율이다. F_1은 정밀도와 재현율의 조화평균으로 F_β에서 β가 1인 경우이다. β의 값이 1보다 큰 경우, 거짓 부정(false negative)에 더 큰 가중치를 줌으로써 정밀도보다 재현율에 더 높은 비중을 둔 조화평균을 계산한다. 특허 문서 분류 모델의 손실 함수에서 클래스 간 불균형을 positive weight β를 통해 고려하였기 때문에, 모델의 평가 지표로써 F_1과 F_β 값을 함께 사용하였다.

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

$$F_1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}$$

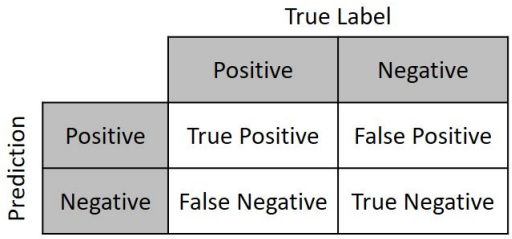

그림 4.2: Confusion Matrix

본 연구에서 제안하는 특허 문서의 다중 레이블 분류 모델은 주어진 특허 문서가 속할 수 있는 IPC 서브클래스 레벨의 모든 레이블에 대해 독립적으로 확률을 예측한다. 그렇기 때문에 하나의 특허 문서에 대해 실제 레이블이 참인 레이블에 비해 거짓인 레이블이 압도적으로 많다. 표 4.1에서 확인할 수 있듯이, USPTO-2M[30] 데이터셋에서 전체 서브클래스 레벨의 IPC 코드는 600개 이상인데 반해, 하나의 특허 문서에는 참값을 갖는 레이블이 평균적으로 2개 미만이다. 따라서 클래스 간 불균형 문제를 해결하기 위한 방법으로 Positive weight β를 하이퍼파라미터로 설정하였다.

4.3 모델의 구성 및 학습

USPTO-2M[30] 데이터셋은 제목과 초록의 텍스트 데이터가 토크나이징된 상태로 제공되기 때문에 기본적인 전처리만 수행하였다. NLTK[33] 자연어처리 패키지를 사용하여 단어들에 대해 불용어(Stopwords)를 제거하고 원형 복원(Lemmatization)을 하였다. 이와 같이 전처리를 거친 단어 토큰을 Gensim[39] 패키지의 Word2vec[34]을 통해 단어 벡터로 임베딩하였다. 단어 임베딩의 Window 크기는 5, 단어 임베딩 차원은 200, 학습률은 0.025로 하였고, 전체 문서에서 등장 횟수가 5번 미만인 단어를 제외한 말뭉치의 크기는 83,002이다.

분류 모델의 학습을 위해 Positive weight β는 positive 레이블의 개수에 대한 negative 레이블의 개수의 비율과 같다고 가정하였다. 이를 검증하기 위해 β의 값을 50, 100, 200, 300, 400으로 달리하여 모델을 학습하였다. GRU 인코더의 은닉층 차원은 50으로 하였고 양방향 GRU 인코더를 사용하기 때문에 문맥 벡터의 차원은 100으로 설정하였다. 학습의 속도를 높이기 위해 단어 벡터 시퀀스를 묶어 크기가 250인 미니배치로 만들어 입력값으로 사용하였다. 이 때 서로 다른 길이의 문서를 동일한 길이(각 미니배치에서 가장 긴 문서의 길이)로 만들기 위해 제로 패딩(Zero padding)을 하였다. 분류 모델의 학습률은 0.01로 설정하였고 Stochastic Gradient Descent를 통해 학습 매개변수를 최적화하였다.

4.4 실험 결과 분석

USPTO-2M[30]을 사용하여 다양한 β의 값에 따라 학습한 모델의 결과를 표 4.2와 그림 4.3, 그림 4.4로 정리하였다. 여기서 Precision, Recall의 값은 백분율이다. 또한 β의 값이 5, 10, 50인 모델의 결과는 50번째 epoch를, 나머지 β의 값이 20, 30, 40, 100, 200, 300, 400인 모델의 결과는 150번째 epoch를 기준으로 정리하였다. 일반적으로 알고리즘의 정밀도와 재현율은 서로 반비례 관계를 가진다. 표 4.2에서도 β가 증가함에 따라 정밀도가 감소하고 재현율이 증가하였다. 이는 3.1.3 절에서 설정한 하이퍼파라미터 β의 역할에 상응하는 결과로써, β의 값이 1보다 큰 경우 특허 문서 분류 모델의 예측값을 참(Positive)으로 강제하는 효과를 보여준다. F_1은 β가 30일 때, F_β는 β가 400일 때 가장 큰 값을 갖는다. 실험 환경의 제약상 더 세밀한 단위의 β에 대한 실험을 하지 못하였지만, 추후에 하이퍼파라미터 조율을 거듭하여 β의 최적해를 찾는다면 더욱 높은 성능의 분류 모델을 기대할 수 있을 것이다.

표 4.2: β에 따른 모델의 결과

β	Precision	Recall	F_1	F_β
5	49.13	7.27	0.1297	0.0772
10	41.61	10.54	0.1723	0.1092
20	27.18	37.41	0.3185	0.3793
30	23.60	50.22	0.3242	0.5081
40	20.65	60.30	0.3107	0.6082
50	14.07	49.00	0.2214	0.4966
100	12.39	80.28	0.2168	0.8057
200	8.74	87.11	0.1602	0.8728
300	7.81	90.12	0.1449	0.9028
400	6.97	91.72	0.1307	0.9184

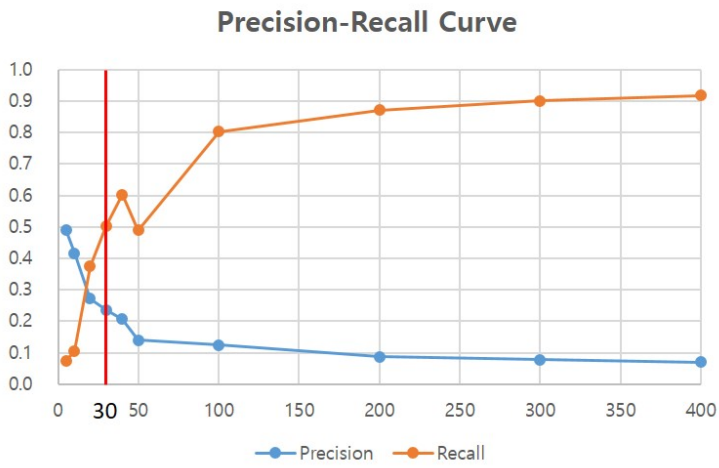

그림 4.3: Precision-Recall Curve

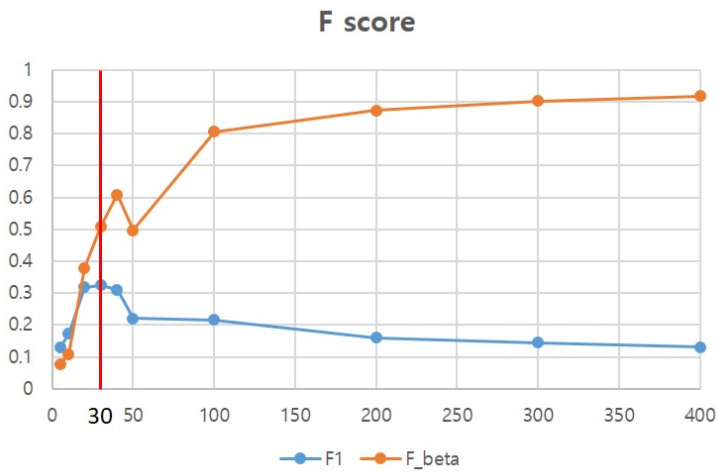

그림 4.4: F_1, F_β Curve

4.5 어텐션 메커니즘을 활용한 키워드 탐색

다음은 특허 문서 분류 모델의 어텐션 메커니즘[5]을 활용하여 특허 문서의 키워드를 탐색하는 방법에 대해 소개한다. 시험 데이터의 특허 문서 중에서 정답을 맞춘 데이터로부터 임의로 5개를 선정하여 그림 4.5, 4.6, 4.7, 4.8, 4.9처럼 나타내었다. 이는 분류 모델에서의 어텐션 스코어를 단어 시퀀스의 순서대로 도식화한 것으로, α_t의 값이 0에 가까우면 어두운 색으로, 1에 가까우면 밝은 색으로 표현하였다. 또한 표 4.3은 각각의 그림에서 나타낸 특허 문서에 대한 세부 정보를 정리한 것이다.

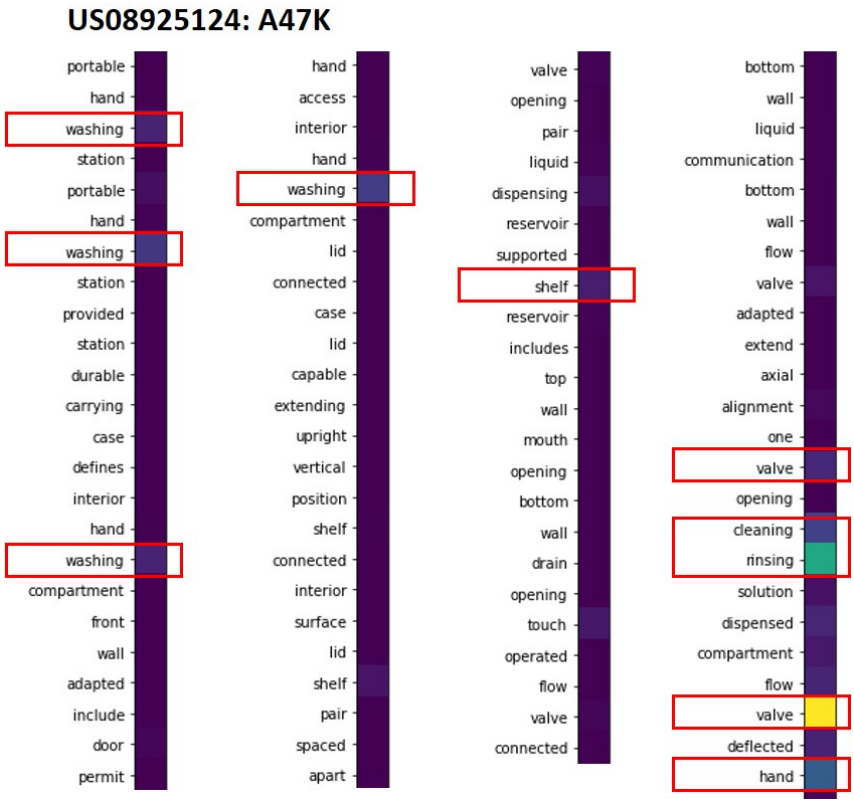

그림 4.5: 특허 문서의 키워드 시각화 (1)

먼저 첫 번째 특허 “US08925124”는 이동식 손 세척기에 대한 발명을 서술하였고 IPC 레이블 A47K에 속한다. A47K 레이블은 Sanitary equipment not otherwise provided for; Toilet accessories, 즉 화장실의 부속품 등과 관련된 위생 설비에 관한 특허를 지칭한다. 그림 4.5의 단어 시퀀스를 살펴보면 ‘washing’, ‘cleaning’, ‘valve’, ‘hand’ 등의 A47K 레이블과 관련된 단어에 하이라이트가 되어있음을 확인할 수 있다.

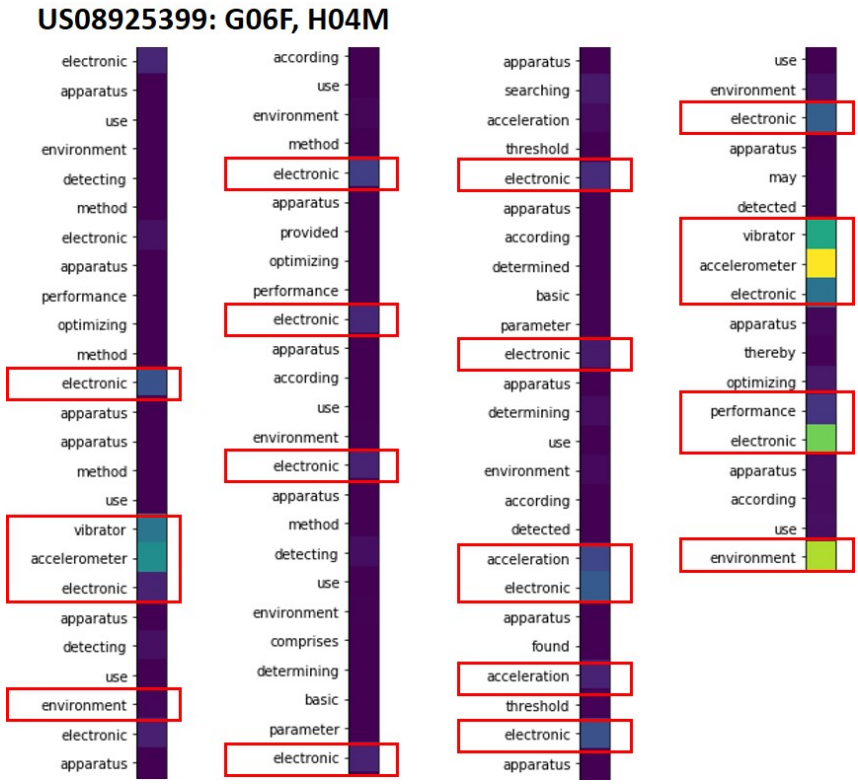

그림 4.6: 특허 문서의 키워드 시각화 (2)

다음 두 번째 특허 “US08925399”는 주변 환경 검출 방법과 관련된 전자 장비의 성능 최적화 기술에 관한 특허로써 IPC 레이블 G06F, H04M에 속한다. G06F 레이블은 Electric digital data processing, 즉 전자 및 디지털 데이터 처리, H04M 레이블은 Telephonic communication, 즉 전화 통신에 관한 특허를 설명한다. 그림 4.6의 단어 시퀀스를 참고하면 ‘electronic’, ‘accelerometer’, ‘acceleration’, ‘environment’와 같이 디지털 데이터 처리와 관련된 단어들과 ‘vibrator’와 같이 전자통신 장비의 부품, 예컨대 스피커에 필요한 진동 장치에 대한 단어의 어텐션 스코어가 높게 나왔다.

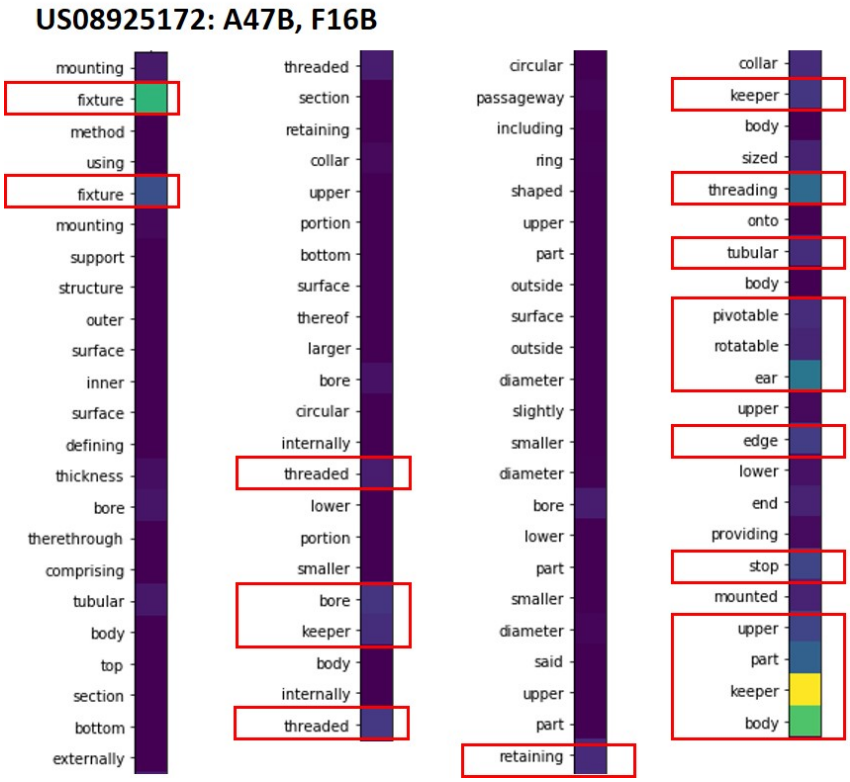

그림 4.7: 특허 문서의 키워드 시각화 (3)

세 번째 특허 “US08925172”는 IPC 레이블 A47B, F16B에 속하는 특허로써 각각 책상, 식탁, 사무용 가구, 캐비닛, 서랍과 같은 일반적인 용도의 가구에 대한 레이블과 못, 볼트, 클립, 경첩 등과 같이 기자재를 연결하고 조이는데 필요한 부품에 관한 레이블을 의미한다. 마찬가지로 그림 4.7의 단어 시퀀스를 살펴보면 ‘tubular’, ‘part’, ‘body’와 같이 레이블 A47B의 주제와 유사한 단어들과 ‘fixture’, ‘threaded’, ‘bore’, ‘retaining’, ‘keeper’와 같이 레이블 F16B와 의미적으로 유사한 단어들의 중요도가 높게 나왔다.

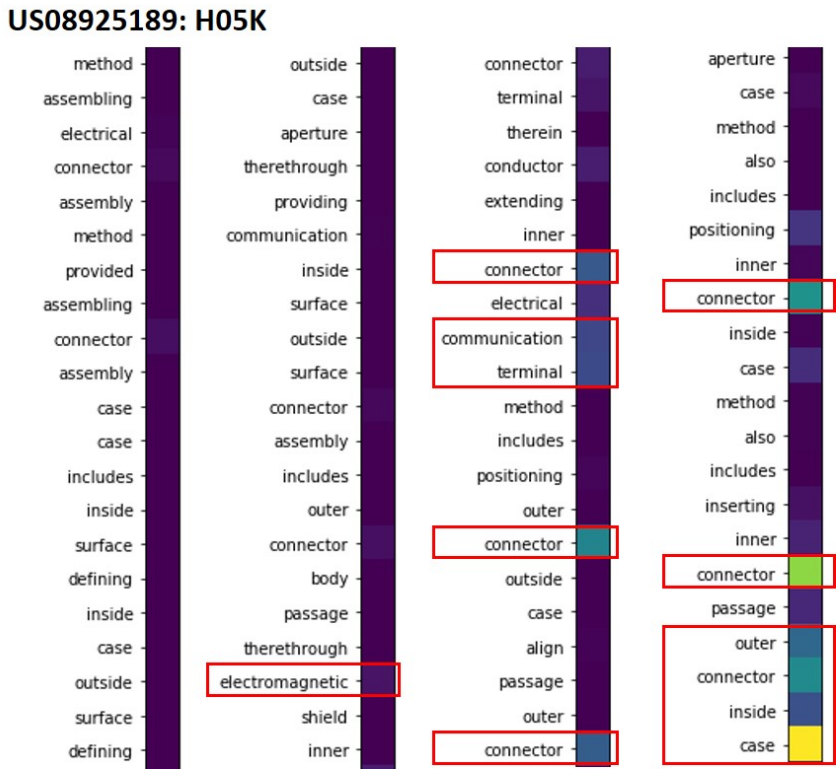

그림 4.8: 특허 문서의 키워드 시각화 (4)

네 번째 특허 “US08925190”은 전자 부품을 탑재하는 작업에 관한 기술을 서술하고 있으며 IPC 레이블 H05K에 속한다. H05K는 인쇄된 회로 기판, 전기 장치의 취급 또는 구조상의 세부 사항, 전기 부품 조립 및 제조에 관한 특허를 포함한다. 그림 4.8의 단어 시퀀스에서도 ‘connector’, ‘electrical’, ‘communication’은 회로 기판과 전기 장치에 관련되어 있고 ‘terminal’, ‘outer’, ‘inside’, ‘case’의 단어는 전기 장치의 구조상의 세부 사항에 관련되어 있음을 확인할 수 있다.

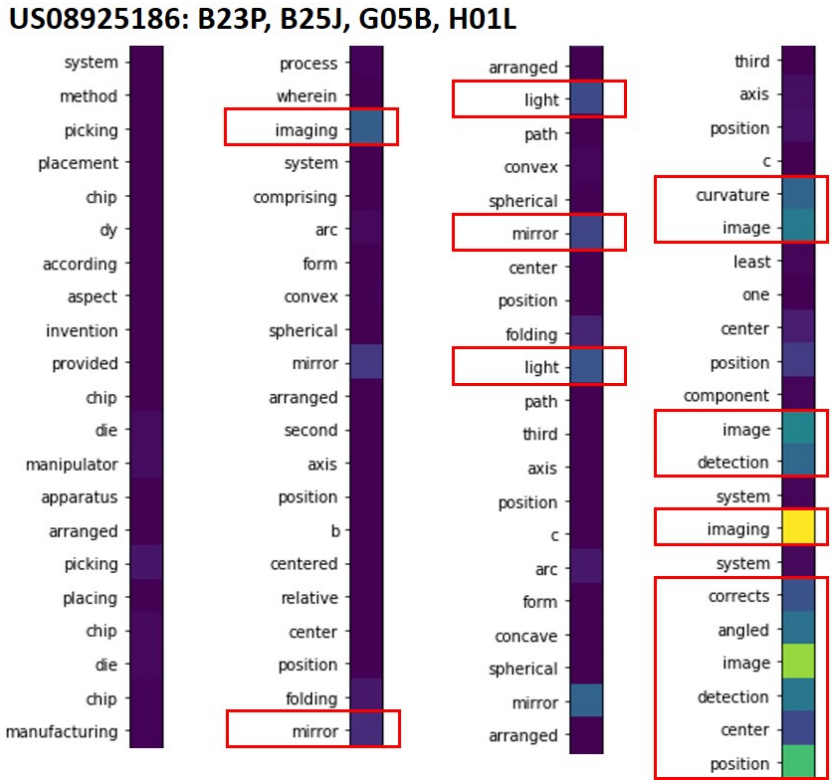

그림 4.9: 특허 문서의 키워드 시각화 (5)

마지막으로 특허 “US08925186”은 IPC 레이블 B23P, B25J, G05B, H01L에 속하는 기술에 관한 것이다. B23P는 기계 공구 및 금속이 제공되지 않는 작업, B25J는 조작자, 조작 장치와 함께 제공되는 챔버, G05B는 일반 시스템 제어 또는 조절, 이러한 시스템의 기능적 요소와 그에 대한 모니터링 또는 테스트, H01L은 반도체 장치에 관한 특허 레이블이다. 그림 4.9의 단어 시퀀스를 확인해보면 ‘imaging’, ‘mirror’, ‘light’, ‘curvature’, ‘detection’, ‘position’ 등과 같이 어텐션 스코어가 높게 나온 단어들이 전반적으로 실제 레이블과 관련된 단어들로 구성됨을 알 수 있다. 따라서 어텐션 메커니즘[5]을 활용한 특허 문서 분류 모델을 통해 각 특허 문서의 주제와 높은 연관성을 지닌 키워드들을 효과적으로 선별할 수 있다.

표 4.3: 그림 4.5, 4.6, 4.7, 4.8, 4.9의 특허 문서 세부 정보

출원번호	IPC	제목
US08925124	A47K	Portable hand washing station
US08925399	G06F, H04M	Electronic apparatus use environment detecting method electronic apparatus performance optimizing method electronic apparatus
US08925172	A47B, F16B	Mounting fixture method using
US08925189	H05K	Electronic component mounting device operation performing method mounting electronic component
US08925186	B23P, B25J, G05B, H01L	System method picking placement

끝으로 특허 문서별 키워드 탐색의 결과 중 일부를 다음의 표 4.4, 4.5, 4.6, 4.7로 정리하였다. 각각의 특허 문서에 대해 t번째 단어의 어텐션 스코어 α_t의 값이 큰 순서대로 단어를 나열하여, 첫 번째부터 10번째까지 단어를 키워드로 선정한 것이다. 이로써 본 연구에서 제안하는 특허 문서 분류 모델을 통해, 새로운 특허 문서가 주어졌을 때 해당 발명이 속할 확률이 높은 레이블의 예측과 함께 문서의 주제를 결정짓는데 영향력이 높은 키워드를 탐색할 수 있다.

표 4.4: 특허 문서별 키워드 탐색 (1)

No	특허번호	실제 IPC 레이블	키워드
1	US08925195	H01R	brush, weapon, lavatory, faucet, oil, grommet, bottle, folding, cleaning, nautical
2	US08925187	H05K	apparel, handrail, fixture, mop, cleaning, portable, gun, cleaning, cleaner, bracket
3	US08925157	B42F	crusher, sweeper, cleaner, brush, curtain, basin, footwear, brick, hair, nautical
4	US08925168	B23P	crib, racquet, tool, hinge, rivet, hinge, dolly, cleaning, theft, gun
5	US08925124	A47K	sweeper, curtain, brush, firing, sipes, faucet, wiping, tie, vehicle, drive
6	US08925255	E04F, E04D, E04B	handrail, decorative, cleaner, display, stowed, door, brush, exchanger, light, press
7	US08925345	F28D, F25B, F25D	crusher, mop, brush, washing, brassiere, button, elastomeric, assistance, cleaning, sheet
8	US08925182	H02K	oil, furniture, cord, toilet, sock, blind, cleaning, locking, nail, child
9	US08925192	H05K, H01K	weapon, hinge, fecal, display, grinding, cleaning, seal, girder, cleaner, robotic
10	US08925399	G06F, H04M	handbag, brush, folding, footwear, metallographic, turbine, sight, human, bracket, hose

표 4.5: 특허 문서별 키워드 탐색 (2)

No	특허번호	실제 IPC 레이블	키워드
11	US08925143	A47L	brush, portable, lifejacket, dirt, tool, ramp, lace, sipe, personnel, seal
12	US08925113	A63B, A41D	faucet, anal, tie, acoustic, dolly, rivet, viewable, hinge, thin, thrust
13	US08925198	B23P, B23B, B24B	crusher, laser, lace, hose, brick, press, turbine, bracket, sanitary, drive
14	US08925121	E04H	defecation, sock, offshore, brush, engine, wheel, butt, valve, preheated, crankcase
15	US08925167	B23P, E03C	racquet, brush, balun, nautical, button, workpiece, oil, suction, bath, indicia
16	US08925172	A47B, F16B	shower, faucet, toilet, ballistic, sipes, dirt, robot, furniture, refrigerant, gun
17	US08925139	A47L	cleaning, toilet, brush, drill, patient, swimming, basin, closure, sterilization, lash
18	US08925196	B23P, F01M, F01L	hair, drilling, dispensing, display, flexible, cleaning, flowable, emptied, accelerometer, bearing
19	US08925117	B32B, A43B, A42B, A41D, F16F	bag, brassiere, cord, cleaning, gun, vest, patient, weight, blade, truck
20	US08925131	E04F	rolling, container, brightness, cleaning, stent, handle, baw, shower, waste, fastener

표 4.6: 특허 문서별 키워드 탐색 (3)

No	특허번호	실제 IPC 레이블	키워드
21	US08925175	B23P, B25B, G02B, B01L	dolly, furniture, workpiece, linen, ball, handle, lubricating, cleaning, pull, cord
22	US08925194	H05K, H01L, H01K	weather, deck, cleaning, tie, magnet, valve, elongate, dielectric, bed, cushioning
23	US08925180	B22F, B23P, B24B, B23Q	sweeper, hydrostatic, firing, cistern, light, ramp, magnet, door, commode, opened
24	US08925186	G05B, B23P, H01L, B25J	paper, hair, tank, lubricating, nautical, cleaning, concussion, acoustic, drive, vessel
25	US08925291	B43M	toilet, washing, sterilization, engaging, mobile, wheel, mortice, substrate, bracket, flux
26	US08925381	G01D, G01C, B63B	shrapnel, lavatory, hand, patient, hose, conductor, caricature, robotic, lace, safety
27	US08925189	H05K	cord, brick, cleaning, grind, door, blade, image, height, oil, stent
28	US08925190	H05K	nail, inferential, roof, girder, decelerating, grip, water, handle, window, speed
29	US08925116	A43B, A41D, A43C	hair, cleat, refrigerant, washing, hinge, tool, fluid, electrode, liquid, bonding
30	US08925170	H05K, B23P, B23K	cleaning, drive, drum, lace, bed, waste, balloon, stringer, ankle, mandrel

표 4.7: 특허 문서별 키워드 탐색 (4)

No	특허번호	실제 IPC 레이블	키워드
31	US08925191	H05K, A61N, H01R	hose, washing, sole, user, drilling, weather, winding, girder, dielectric, baw
32	US08925120	A47K	grind, cleaning, workpiece, laser, magnet, cleat, abdomen, filter, solder, cantilever
33	US08925159	E06B	offshore, bag, eye, wearer, tenon, exchange, shower, clamping, filter, bearing
34	US08925176	B23Q, B23P, F16L, F24F, F25B	wiping, washing, camera, cleaning, lace, handle, sterilization, strap, unlock, clip
35	US08925197	B23P, F01D	patient, nautical, dirt, cleaning, valve, picking, decelerating, waste, drum, wheel
36	US08925122	A47K	sleeping, conductor, roof, cable, drill, refrigerant, bearing, strut, shim, intake
37	US08925129	A43B, A43C	brassiere, ankle, climbing, book, sole, sterilization, pouch, basin, cleaned, patterning
38	US08925138	A47L	turbine, child, stent, lace, oil, blind, tongue, hinge, bondable, spring
39	US08925173	B23P, B27F	brush, stair, emptying, camera, sponge, flipping, upstream, surge, sterilize, bracket
40	US08925193	H05K, G01R, H01K	plant, handbag, oil, rafter, electrode, drive, stepping, pilot, concussion, shower

제 5 장 결론

최근 지적재산권과 관련한 특허의 출원은 지속적으로 증가하는 추세이다. 하지만 특허 심사는 여전히 수작업에 강하게 의존하는 경향이 있어, 특허 등록이 기술의 발전 속도를 따라잡지 못하고 있다. 이러한 상황에서 본 연구가 제안하는 특허 문서의 다중 레이블 분류에 관한 방법론은 다음과 같은 의의를 갖는다. 첫 번째, 대량의 특허 정보를 주제에 따라 효과적으로 분류하여 향후 특허 심사에 소요되는 시간과 비용을 획기적으로 줄일 수 있을 것으로 기대한다. 특히 최소한의 전처리를 통해 2백만 건에 가까운 특허 문서의 모든 단어에 대해 자동적으로 피처를 추출할 수 있다는 것이 기존 연구와의 차별점이다. 두 번째, 어텐션 메커니즘[5]을 통해 특허 문서의 주제를 결정짓는 키워드를 탐색할 수 있다. 특히 기존 특허 분류 모델[30]과 달리 주어진 문서의 길이에 상관없이 특허 문서의 중요한 정보를 효과적으로 피처 벡터로 추출할 수 있다. 이를 통해 특허 문서의 분류 결과에 대한 단어 레벨의 해석을 제공하여 향후 특허 심사를 보조하는 역할로써 활용될 수 있을 것으로 기대한다. 세 번째, 본 연구가 제안하는 방법은 특허 문서를 IPC의 서브클래스 레벨뿐만 아니라 목적에 따라 다양한 종류와 레벨의 기준으로 분류할 수 있다. IPC뿐만 아니라 미국특허분류 USPC 또는 CPC (Cooperative Patent Classification)와 같은 분류기준에 대해서도 쉽게 레이블만 바꿔 학습할 수 있어 범용적으로 활용할 수 있다.

본 연구의 한계점도 존재한다. 먼저, IPC 레이블은 본 연구에서 중점적으로 다룬 서브클래스 레벨(약 640개)보다 세분화된 서브그룹 레벨(약 72,000개)까지 구성된다. 그렇기 때문에 제안하는 특허 문서 분류 모델을 더 세분화된 단위의 IPC 레이블에 대해 모델링을 할 필요가 있다. 또한 특허 문서는 일반적인 자연어와 달리, 특수 분야에서만

사용되는 전문적인 용어(jargon)의 사용이 빈번하다. 이러한 용어들은 해당 문서에서는 중요한 의미를 지니지만 전체 문서 대비 출현 빈도가 낮다. 그렇기 때문에 단어 벡터 학습을 할 때, 전문적 용어를 모두 반영한 단어 사전을 만드는 것에 다소 어려움이 있다.

향후 연구 과제로는 단어와 문장, 문서의 계층적인 구조를 고려한 문서 분류 모델[51]을 적용하여 더 정교한 다중 레이블 특허 분류를 해보고자 한다. 특히 예측 레이블에 대해 독립적으로 어텐션 메커니즘[5]을 적용한다면 보다 정교한 키워드 탐색이 가능할 것으로 기대한다. 또한 본 연구에서 특허 문서의 피처 벡터를 얻기 위해 사용한 특허 정보의 활용 범위를 넓히는 것도 좋은 접근 방법이다. 특허 문서의 제목과 초록 항목보다 길이가 길고 발명의 구체적인 정보가 포함된 명세서와 청구항의 텍스트 데이터, 그리고 다른 특허 문서들과의 상관 관계를 나타내는 발명권자, 국내외 참조특허와 같은 서지정보를 함께 활용한다면 더욱 정확한 특허 문서 분류를 할 수 있을 것이다.

참고 문헌

[1] *Patentsview*. http://www.patentsview.org/download/.

[2] *Uspto patent application full-text and image database*. http://appft.uspto.gov/netahtml/PTO/search-bool.html.

[3] *특허정보넷 키프리스*. http://www.kipris.or.kr/khome/main.jsp#.

[4] *Guide to the international patent classification version 2018*, 2018.

[5] D. Bahdanau, K. Cho, and Y. Bengio, *Neural machine translation by jointly learning to align and translate*, CoRR, abs/1409.0473 (2014).

[6] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, *A neural probabilistic language model*, J. Mach. Learn. Res., 3 (2003), pp. 1137–1155.

[7] K. Benzineb and J. Guyot, *Automated patent classification*, (2011).

[8] Y.-L. Chen and Y.-C. Chang, *A three-phase method for patent classification*, Information Processing Management, 48 (2012), pp. 1017 – 1030.

[9] J. Chiu and E. Nichols, *Named entity recognition with bidirectional lstm-cnns*, Transactions of the Association for Computational Linguistics, 4 (2016), pp. 357–370.

[10] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, *Learning phrase representations using RNN*

encoder-decoder for statistical machine translation, CoRR, abs/1406.1078 (2014).

[11] R. Collobert and J. Weston, *A unified architecture for natural language processing: Deep neural networks with multitask learning*, in Proceedings of the 25th International Conference on Machine Learning, ICML '08, New York, NY, USA, 2008, ACM, pp. 160–167.

[12] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, *Natural language processing (almost) from scratch*, J. Mach. Learn. Res., 12 (2011), pp. 2493–2537.

[13] E. D'hondt, S. Verberne, N. Weber, K. Koster, and L. Boves, *Using skipgrams and pos-based feature selection for patent classification*, Computational Linguistics in the Netherlands Journal, (2012), pp. 52–70.

[14] C. J. Fall, A. Törcsvári, K. Benzineb, and G. Karetka, *Automated categorization in the international patent classification*, SIGIR Forum, 37 (2003), pp. 10–25.

[15] X. Guo, L. Gao, X. Liu, and J. Yin, *Improved deep embedded clustering with local structure preservation*, in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 1753–1759.

[16] J. Guyot, K. Benzineb, and G. Falquet, *myclass: A mature tool for patent classification*, in CLEF, 2010.

[17] K. M. Hermann, T. Kociský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, *Teaching machines to read and comprehend*, CoRR, abs/1506.03340 (2015).

[18] R. Johnson and T. Zhang, *Effective use of word order for text categorization with convolutional neural networks*, CoRR, abs/1412.1058 (2014).

[19] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, *A convolutional neural network for modelling sentences*, CoRR, abs/1404.2188 (2014).

[20] Y. Kim, *Convolutional neural networks for sentence classification*, in EMNLP, 2014.

[21] Y. G. Kim, J. H. Suh, and S. C. Park, *Visualization of patent analysis for emerging technology*, Expert Syst. Appl., 34 (2008), pp. 1804–1812.

[22] A. Kumar, O. Irsoy, J. Su, J. Bradbury, R. English, B. Pierce, P. Ondruska, I. Gulrajani, and R. Socher, *Ask me anything: Dynamic memory networks for natural language processing*, CoRR, abs/1506.07285 (2015).

[23] S. Lai, L. Xu, K. Liu, and J. Zhao, *Recurrent convolutional neural networks for text classification*, in Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, AAAI'15, AAAI Press, 2015, pp. 2267–2273.

[24] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, and C. Dyer, *Neural architectures for named entity recognition*, in Proceedings of the 2016 Conference of the North American Chapter of the Association

for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2016, pp. 260–270.

[25] L. S. Larkey, *A patent search and classification system*, in Proceedings of the Fourth ACM Conference on Digital Libraries, DL '99, New York, NY, USA, 1999, ACM, pp. 179–187.

[26] Q. V. Le and T. Mikolov, *Distributed representations of sentences and documents*, CoRR, abs/1405.4053 (2014).

[27] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*, Proceedings of the IEEE, 86 (1998), pp. 2278–2324.

[28] J. Leskovec, J. Kleinberg, and C. Faloutsos, *Graphs over time: Densification laws, shrinking diameters and possible explanations*, in Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, KDD '05, New York, NY, USA, 2005, ACM, pp. 177–187.

[29] J. Li, T. Luong, and D. Jurafsky, *A hierarchical neural autoencoder for paragraphs and documents*, in Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Association for Computational Linguistics, 2015, pp. 1106–1115.

[30] S. Li, J. Hu, Y. Cui, and J. Hu, *Deeppatent: patent classification with convolutional neural networks and word embedding*, Scientometrics, (2018), pp. 1–24.

[31] Z. Li, D. Tate, C. Lane, and C. Adams, *A framework for automatic triz level of invention estimation of patents using natural language processing, knowledge-transfer and patent citation metrics*, Comput. Aided Des., 44 (2012), pp. 987–1010.

[32] R. Lin, S. Liu, M. Yang, M. Li, M. Zhou, and S. Li, *Hierarchical recurrent neural network for document modeling*, in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2015, pp. 899–907.

[33] E. Loper and S. Bird, *Nltk: The natural language toolkit*, in Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1, ETMTNLP '02, Stroudsburg, PA, USA, 2002, Association for Computational Linguistics, pp. 63–70.

[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, CoRR, abs/1301.3781 (2013).

[35] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, *Distributed representations of words and phrases and their compositionality*, CoRR, abs/1310.4546 (2013).

[36] T. Mikolov, S. W.-t. Yih, and G. Zweig, *Linguistic regularities in continuous space word representations*, in Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics:

Human Language Technologies (NAACL-HLT-2013), Association for Computational Linguistics, May 2013.

[37] J. Pennington, R. Socher, and C. D. Manning, *Glove: Global vectors for word representation.*, in EMNLP, vol. 14, 2014, pp. 1532–1543.

[38] F. Piroi, M. Lupu, A. Hanbury, and V. Zenz, *Clef-ip 2011: Retrieval in the intellectual property domain*, in CLEF, 2011.

[39] R. Řehůřek and P. Sojka, *Software Framework for Topic Modelling with Large Corpora*, in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, Valletta, Malta, May 2010, ELRA, pp. 45–50. `http://is.muni.cz/publication/884893/en`.

[40] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, *Neurocomputing: Foundations of research*, MIT Press, Cambridge, MA, USA, 1988, ch. Learning Representations by Back-propagating Errors, pp. 696–699.

[41] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts, *Recursive deep models for semantic compositionality over a sentiment treebank*, in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2013, pp. 1631–1642.

[42] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, *Weakly supervised memory networks*, CoRR, abs/1503.08895 (2015).

[43] K. S. Tai, R. Socher, and C. D. Manning, *Improved semantic representations from tree-structured long short-term memory networks*, CoRR, abs/1503.00075 (2015).

[44] D. Tang, B. Qin, and T. Liu, *Document modeling with gated recurrent neural network for sentiment classification*, in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2015, pp. 1422–1432.

[45] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, and G. E. Hinton, *Grammar as a foreign language*, CoRR, abs/1412.7449 (2014).

[46] I. Vulić and M.-F. Moens, *Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings*, in Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15, New York, NY, USA, 2015, ACM, pp. 363–372.

[47] Wipo, *World Intellectual Property Indicators, 2017 edition*, World Intellectual Property Organization - Economics and Statistics Division, 2017.

[48] C.-H. Wu, Y. Ken, and T. Huang, *Patent classification system using a new hybrid genetic algorithm support vector machine*, Appl. Soft Comput., 10 (2010), pp. 1164–1177.

[49] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, *Show, attend and tell: Neural image caption generation with visual attention*, in Proceedings of the 32Nd Interna-

tional Conference on International Conference on Machine Learning - Volume 37, ICML'15, JMLR.org, 2015, pp. 2048–2057.

[50] Z. Yang, X. He, J. Gao, L. Deng, and A. J. Smola, *Stacked attention networks for image question answering*, CoRR, abs/1511.02274 (2015).

[51] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy, *Hierarchical attention networks for document classification*, in HLT-NAACL, 2016.

[52] D. Yogatama, M. Faruqui, C. Dyer, and N. A. Smith, *Learning word representations with hierarchical sparse coding.*, in ICML, F. R. Bach and D. M. Blei, eds., vol. 37 of JMLR Proceedings, JMLR.org, 2015, pp. 87–96.

[53] T. Young, D. Hazarika, S. Poria, and E. Cambria, *Recent trends in deep learning based natural language processing*, CoRR, abs/1708.02709 (2017).

[54] X. Zhang, J. J. Zhao, and Y. LeCun, *Character-level convolutional networks for text classification*, CoRR, abs/1509.01626 (2015).

[55] C. Zhou, C. Sun, Z. Liu, and F. C. M. Lau, *A C-LSTM neural network for text classification*, CoRR, abs/1511.08630 (2015).

Abstract

Multi-label Patent Classification with Attention Mechanism

Nohil Park
Department of Industrial Engineering
The Graduate School
Seoul National University

Recently, the growth of number of patent application is unprecedented globally. Meanwhile the patent examination is still strongly dependent on manual works by few patent experts, which slows the overall patent registration process. Therefore, an automatic patent classification algorithm is necessary. In this paper, we propose an effective multi-label patent classification algorithm based on the GRU encoder and attention mechanism. We use the USPTO-2M data set, which consists of about 2 million US patent documents, to train our patent classification model. Precision, recall, F_1 score, and F_β score are used to evaluate our model on multi-label patent classification task. By visualizing the attention scores, we could identify and analyze keywords from each patent document which determine the context and IPC codes for subclass level.

Keywords: Patent Classification, Multi-label Classification, Attention Mechanism,

Gated Recurrent Unit, Document Encoder, Word Embedding

Student Number: 2017-22964