석사학위논문

시퀀스 생성 모델을 활용한 특허 문서의 계층적 다중 레이블 분류

Hierarchical multi-label classification of patent documents using sequence generation model

윤 승 주

한양대학교 대학원

2021년 2월

석사학위논문

시퀀스 생성 모델을 활용한 특허 문서의 계층적 다중 레이블 분류

Hierarchical multi-label classification of patent documents using sequence generation model

지도교수 김종우

이 논문을 경영학 석사학위논문으로 제출합니다.

2021년 2월

한양대학교 대학원

비즈니스인포매틱스학과

윤 승 주

이 논문을 윤승주의 석사학위 논문으로 인준함

2021년 2월

심 사 위 원 장 : 장 석 권 (인)

심 사 위 원 : 임 규 건

심 사 위 원 : 김 종 우 (인)

한양대학교 대학원

차 례

표 차 례

그 림 차 례

국 문 요 지

현실에서 수집할 수 있는 문서들은 계층적인 구조를 이루고 있는 경우가 많다. 예로 들어 전자 문서들을 살펴보면 문서들은 다양한 카테고리로 구분이 가능하고 각각의 카테고리는 여러 개의 하위 계층의 카테고리로 세분화 할 수 있다. 특히 특허 데이터의 경우 가장 계층이 낮은 레이블을 기준으로 볼 때 문서를 분류할 수 있는 레이블 수는 굉장히 많으며 보다 정확한 분류를 하기 위해 다양한 방법의 연구가 진행되고 있다.

기존 연구에서는 특허 문서 분류에 대해서 멀티 레이블 분류 문제를 중심으로 연구가 진행되어 왔다. 하지만 이러한 연구 방법들은 레이블 간의 의존성을 무시하며 IPC (International Patent Classification) 레이블의 계층적 구조인 점을 충분히 적용하지 못하고 있다.

본 연구에서는 이러한 한계점을 극복하기 위해 기계번역에 좋은 성능을 보인 트랜스포머(Transformers)를 활용하여 계층적 멀티 레이블 분류 모델을 소개한다. 특허의 요약 문서를 활용하여 IPC 레이블을 섹션과 클래스, 서브클래스로 분류를 하였으며 기존에 특허 분류 연구에서 부족했던 멀티 레이블 간의 관계를 포함할 수 있었다.

본 연구는 특허 분류에 대한 성능 평가를 위하여 USPTO-2M 데이터를 사용했으며 최종적으로 특허 문서 분류 평가 지표인 Top Prediction, Three Guesses, All Categories 중에서 Top Prediction과 All Categories에서 기존 모델보다 분류 성능이 뛰어나다는 것을 확인했다.

Keyword 특허 분류, 생성 모델, 기계 번역, 계층적 분류, 멀티 레이블 분류

제1장 서론

제1절 연구 배경

현실에서 수집할 수 있는 문서들은 계층적인 구조를 이루고 있는 경우가 많으며 각각의 레이블이 하위 계층의 레이블을 가지는 경우를 볼 수 있다. 예로 들어 웹 페이지나 특허 문서, 혹은 이메일 같은 전자 문서들을 살펴보면 문서들은 다양한 카테고리로 구분이 가능하고 각각의 카테고리는 여러 개의 하위 계층의 카테고리로 세분화 할 수 있다. 특히 특허 데이터의 경우 가장 계층이 낮은 레이블을 기준으로 볼 때 문서를 분류할 수 있는 레이블 수는 굉장히 많으며 보다 정확한 분류를 하기 위해 다양한 방법의 연구가 진행되고 있다.

특허 분류를 하는데 있어서 많은 연구들이 멀티 레이블 분류 문제로 바라보고 많은 연구가 진행되어 왔다. 하지만 IPC 레이블을 보다 복잡한 구조를 가지고 있다. 계층적이며 각 계층마다 여러 개의 클래스가 존재하며 특허 문서는 하나가 아닌 두 개 이상인 레이블을 가지는 멀티 레이블 문제이다.

본 연구에서는 정제된 데이터가 아닌 실제 데이터의 텍스트를 활용하여 IPC 분류 코드를 시퀀스 생성 모델을 통해 분류하는 문제를 다루고자 한다. 연구 방향으로는 기계번역을 중심으로 활용되고 있는 모델을 계층적 멀티 레이블 분류에 적용해볼 것이며 편향성이 정제된 데이터가 아닌 현실 데이터를 활용하여 모델을 구현할 것이다. 마지막으로 레이블 간의 의존성을 적용할 수 있는 모델을 구현하는 방향으로 연구를 진행한다.

제2절 논문 구성

본 논문의 구성은 다음과 같다. 1장 서론에 이어서 2장에서는 IPC 분류와 계층적 분류 방법에 대한 배경 지식을 소개하고 시퀀스 생성 모델에 대해 서술한다. 3장에서는 시퀀스 모델을 활용하여 특허 문서 분류 모델을 제시하고, 4장에서는 평가 데이터 셋 및 평가 지표를 소개하며 제안한 모델의 성능을 평가한다. 마지막으로 5장에서는 본 연구의 결론 및 향후 연구를 제시한다.

제2장 관련 연구

제1절 IPC 분류

International Patent Classification(IPC)은 특허 문서를 분류하기 위한 계층적 분류 시스템으로 섹션, 클래스, 서브클래스, 그룹, 그리고 서브그룹으로 나눠진다. World Intellectual Property Organization (WIPO)에서 발표한 Guide to the International Patent Classification(2020)에 따르면 8개의 섹션과 130개의 클래스, 640개의 서브클래스, 그리고 7,400개의 그룹과 72,000개의 서브그룹으로 구성되어 있다. 각각의 계층들은 영어 알파벳과 숫자의 조합으로 분류가 되며 기술 분야를 8개로 나눈 섹션 계층은 <표 1>와 같이 알파벳 대문자 A부터 H로 지정했다. <표 2>는 IPC 분류의 예시를 보여준다. 섹션 F를 예를 들면, 모든 기계 공학과 조명, 가열, 무기, 폭파와 관련된 특허 문서들은 섹션 F로 분류가 된다. 그 중에서 연소 기관과 관련된 문서는 섹션 아래 클래스

<표 1> IPC 섹션 별 분야

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

중에서 F02로 분류되며, 연소 기관의 제어와 관련된 문서는 F02D, 가연성 혼합물 혹은 구성 성분의 전기적 제어는 그룹 F02D 41, 마지막으로 제어신호를 발생하는 회로장치는 서브그룹 F02D 41/02로 분류되는 것을 볼 수 있다.

<표 2> IPC 분류의 예시

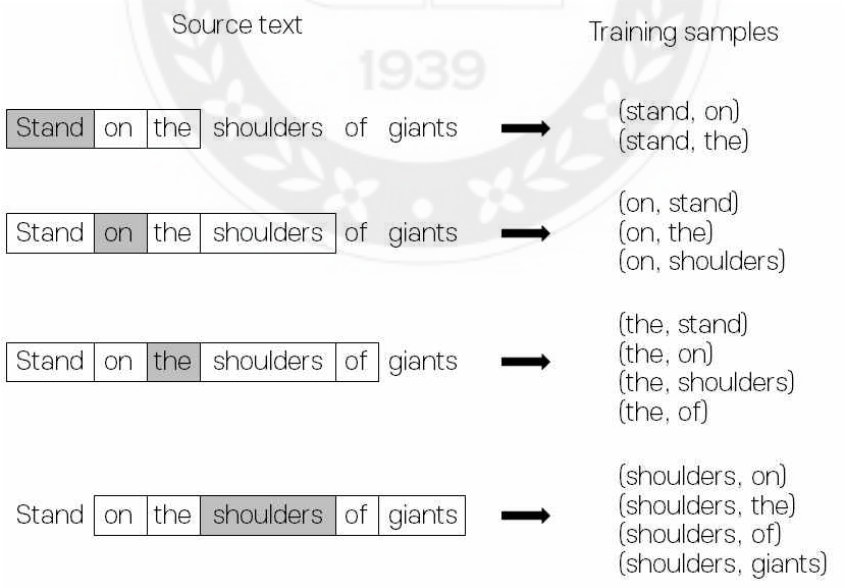
레이블 계층	IPC 레이블	분야
섹션	F	기계 공학, 조명, 가열, 무기, 폭파
클래스	F02	연소기관, 열가스 또는 연소생성물을 이용하는 기관설비
서브클래스	F02D	연소 기관의 제어
그룹	F02D 41	가연성 혼합물 또는 그 구성 성분의 전기적 제어
서브그룹	F02D 41/02	제어신호를 발생하는 회로장치

또한, 특허에 부여되는 IPC 레이블은 Main Category와 Incidental Category로 나뉜다. 모든 특허 문서는 하나의 Main Category에 해당하는 IPC 레이블로 분류되며 Main Category 외 추가로 분류되는 IPC 레이블은 Incidental Category로 간주된다(Fall et al, 2003).

제2절 단어 임베딩

단어 임베딩이란 텍스트를 컴퓨터가 이해할 수 있는 숫자로 표현하는 방법으로 자연어 말뭉치를 통해 각각의 단어들의 의미를 벡터 공간에 표현하는 것을 말한다. 단어 임베딩 방법으로는 대표적으로 원-핫-인코딩(one-hot-enocding)과 워드투벡터(word2vector)가 있다.

원한인코딩이란 텍스트 데이터의 단어 사전 크기만큼의 영벡터에서 각 단어에 해당하는 원소만을 1로 표현하는 방식으로 희소표현이라고도 한다. 이러한 단어 표현 방식은 가장 단순한 방법이지만 어휘 사전의 크기가 커질수록 차원이 커지기 때문에 차원의 저주에에 빠질 수 있으며 단어 간의 유사도를 계산할 수 없다는 한계점이 존재한다.(김영수, 이승우, 2018, 최윤수, 2018)



<그림 1> 윈도우 크기가 2일 때의 학습 데이터

이러한 한계점을 극복하고자 나온 모델이 워드투벡터이다. 워드투벡터는 분포 가설이라는 가정 하에 단어를 벡터로 표현하는 방법이다. 즉, 비슷한 위치에 등장하는 단어들은 비슷한 의미를 가진다는 가정을 가지고 모델 학습이 진행된다. 단어를 벡터 공간에 표현하는 방식으로 단어 간의 유사도를 계산할 수 있게 되었다(Mikolov et al, 2013).

워드투벡터 모델을 구현하는데 있어서 두 가지 방법이 있다. Continuous Bag of Words(CBOW)는 주변 단어 n개를 통해서 중앙에 위치한 단어를 예측하는 방법이다. 예를 들어 'Stand on the shoulders of giants'라는 문장을 예시로 입력과 출력을 살펴보면 다음과 같다. 윈도우 사이즈 크기가 2인 경우 <그림 1>과 같이 학습 데이터가 생성된다. Skip gram은 cbow와 반대로 중앙에 있는 단어를 통해서 주변 단어를 예측하는 방법이다. CBOW 방법은 Skip-gram 보다 학습 속도가 빠르며 학습 데이터가 상대적으로 적어도 활용이 가능한 반면 Skip-gram은 CBOW 보다 학습 기간에 있어서 오래 걸리지만 대량의 데이터에 유용하다는 장점이 있다.

제3절 특허 문서 분류

특허 문서 분류에 대한 연구는 분류 모델로 잘 알려진 SVM과, k-NN, 그리고 인공신경망 등을 활용하여 연구가 진행되어 왔다. Naive Bayes, SVM, k-NN 모델을 활용하여 진행한 연구에서는 서브클래스까지 분류된 IPC 특허 문서 데이터를 사용했으며 제목과 요약문, 그리고 청구항 항목을 활용하여 특허 문서를 분류하였다. 분류 결과 평가 방법 중에서 Top Prediction과 All Categories에서 SVM이 가장 좋은 결과를 보여줬으며 k-NN이

Three-Guesses에서 가장 좋은 결과를 보여줬다(Fall et al., 2003). k-NN을 활용한 또 다른 연구로는 특허의 제목과 요약문, 배경, 초록을 활용하여 미국특허분류 코드인 United States Patent Classification (USPC)를 분류하였으며 텍스트 문장과 청구항 항목을 사용했을 때 분류 모델의 성능이 가장 좋은 결과를 보였다(Larkey et al., 1999).

특허 데이터의 일부를 활용해서 진행한 연구도 있었다. 반도체 장비와 관련된 특허 문서로부터 키워드를 추출한 뒤 유전 알고리즘을 기반으로 한 SVM 모델과 USPC 데이터 중에서 360/324에 해당하는 특허 데이터를 사용하여 2층 순전파 인공신경망과 Levenberg-Marquardt 알고리즘을 활용한 분류 모델을 제안하였다. 하지만 이러한 연구의 한계점으로는 연구에 사용된 데이터 수가 적어 실제 대량의 데이터가 주어졌을 때 모델에 적용하기에는 한계가 있다는 점이다(Wu et al., 2010; Guyot et al., 2010 Li et al., 2012).

최근에는 딥러닝을 활용한 연구가 활발히 진행 중에 있다. 단어 임베딩과 합성곱 신경망을 활용하여 임베딩된 단어를 3가지 필터를 통해 특징을 추출했으며 IPC 레이블에서 서브클래스 레벨까지에 대한 특허 분류 모델을 제안했다(Li et al 2018). 또한 순환 신경망을 활용해서 USPTO 벌크 데이터 2006년에서 2017년까지의 특허 데이터를 분류하는 모델을 제안했다. 하지만 여기서 사용한 특허 데이터는 총 50개의 클래스로 제한되어 있었다(Grawe et al., 2017).

제4절 계층적 분류

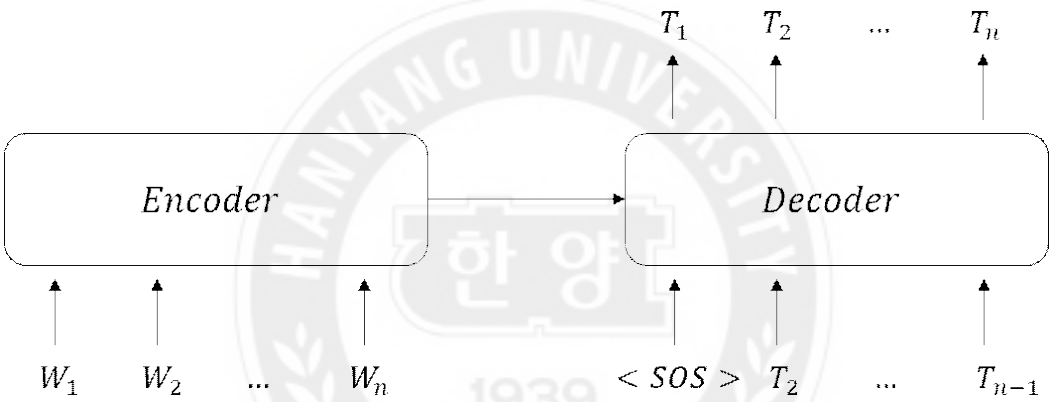
계층적 분류를 접근하는 방법은 세 가지가 있다. Flat 접근법은 모든 계층 레벨을 고려하여 발생할 수 있는 모든 레이블을 구성한 후 멀티 레이블로 접근하는 법이다. 즉, 가장 마지막 계층을 레이블로 설정하여 접근하는 간단한 접근 방법으로 Naive Bays와 같은 분류 모델을 사용한다(Fall et al, 2003). 하지만 이런 접근 방식은 레이블의 계층 정보를 무시한다.

이러한 문제점을 해결하기 위해서 선행 연구에서는 다음과 같은 두 가지 방법을 제시했다. 우선 Local 접근 방법은 하향식 접근 방법으로 각 계층, 혹은 해당되는 카테고리마다 분류 모델을 구현하는 방법이다. 각 분류 모델들은 서로 정보를 공유하지 않는다라는 문제점이 있다. 하지만 위에 제시안 Flat 접근 방법과 Local 접근 방법은 상위 계층의 분류 오류를 피할 수 있다는 장점이 잇다. Global 접근 방법은 하나의 분류 모델만을 사용하여 계층 분류를 하며 기계학습과 딥러닝 방법을 통해 연구가 진행되어 왔다(Chen et al, 2004, Xue et al, 2008, Peng et al, 2018, Yan et al 2016) 기존 연구들은 계층 분류를 위해 모델 구조를 변형하거나 손실함수를 만들었다.

하지만 Global 접근 또한 계층적 분류 관계를 충분히 활용하지 않고 있다. 문제점은 모델이 분류를 진행하는데 있어서 동일한 문서 표현 방식을 사용한다는 것이다. 즉, 한 번 임베딩한 것을 모든 계층 분류를 하는데 사용했으며 모델의 출력값을 다음 계층을 예측한는데 사용되지 않았다는 문제점이 있다.

제5절 시퀀스 생성 모델

시퀀스 투 시퀀스 모델은 하나의 시퀀스를 다른 하나의 시퀀스로 변환하는 자연어처리 기반의 딥러닝 모델이며 기계 번역 및 대화 생성을 기점으로 다양한 분야에서 연구가 진행되고 있다.



<그림 2> 시퀀스 투 시퀀스 모델의 구조

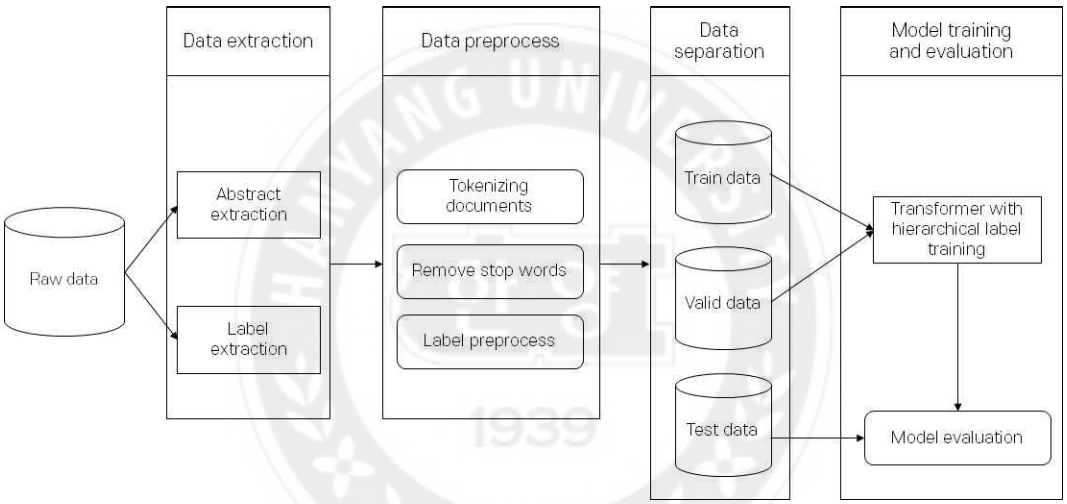
시퀀스 투 시퀀스 모델의 구조는 크게 인코더와 디코더로 나뉘며 인코더는 입력 시퀀스를 받아서 입력 받은 데이터의 특징을 추출하는 역할을 하고, 디코더는 t시점에서 추출된 특징 값을 입력 받아 t+1 시점의 시퀀스를 출력하는 구조로 되어 있다(<그림 2> 참조). Bahdanau et al.(2014)에서 제안한 모델은 시퀀스 투 시퀀스 모델과 어텐션 개념를 함께 사용하여 기계번역의 성능을 높이는데 기여했으며 인공신경망에서 블랙 박스로 여겨졌던 부분을 시각화할 수 있었다. 또한 Vaswani et al.(2017)에서는 기존 연구에서 순환 신경망을 기반

으로 한 시퀀스 투 시퀀스 모델의 한계점을 극복하였으며 임베딩 된 단의들의 특징을 더 정교하세 추출하고 모델 학습 시간 또한 줄이며 성능 또한 높였다.

하지만 시퀀스 모델 또한 단점이 존재한다. LSTM 기반으로 이루어진 모델은 학습 시간이 길다. 이를 해결하기 위해 LSTM을 사용하지 않으면서 시퀀스적인 의미를 담을 수 있도록 구현된 모델이 트랜스포머 모델이다. 트랜스포머 모델은 구글에서 발표를 했으며 기계번역에 있어서 훌륭한 성능을 보였다.

3. 제안 모델

본 연구에서는 트랜스포머 모델을 바탕으로 특허 문서의 계층적 멀티 레이블 분류 모델을 제안하고자 한다. 특허 문서 분류 모델의 전체적인 연구 절차는 다음과 같다(<그림 3> 참조).



<그림 3> 제안하는 방법의 전체적인 연구 모형

연구 절차는 크게 4가지로 나뉜다. 우선, 특허 데이터로부터 특허 문서의 요약문과 IPC 레이블을 추출해낸다. 추출된 데이터는 불용어 처리와 토크나이징, 그리고 레이블 전처리 과정을 거친 후 모델 학습과 검정, 평가 데이터로 나눈다. 학습 데이터와 검증 데이터를 통해 모델 학습을 마치면 테스트 데이터로 모델을 평가하는 방법으로 연구를 진행한다. 최종적으로 모델은 새로운 특허 문서가 주어졌을 때 특허의 요약문을 활용하여 서브클래스 계층의 IPC 레이블을 분류하는 것을 목표로 한다.

제1절 특허 문서 분류를 위한 데이터 추출 및 전처리

본 연구에서는 특허 문서의 요약문을 통해 IPC 레이블 분류를 위한 모델을 제안한다. 모델을 설명하기 앞서 데이터 추출과 전처리 과정을 소개한다. 우선, 가공되지 않은 현실 데이터에서 특허 문서의 요약문과 IPC 레이블을 추출한 후에 데이터 가공을 진행하며 진행 과정은 아래 <표 3>과 같이 진행하였다.

우선 불용어 처리 과정을 통해서 문서를 임베딩하는데 불필요한 단어들을 제거했다. 이후 Spacy 패키지를 통해 문서를 단어 단위로 토큰화를 했으며 표제어 추출 또한 진행을 하였다. 또한 학습 데이터를 통해서 단어 사전을 구축했으며 <UNK>, <PAD> 토큰도 포함을 시켰다. 학습 데이터에서 구축된 단

<표 3> 데이터 전처리 진행

요약문	원본 데이터	a hockey helmet for receiving a head of a wearer the head having a crown region...
	불용어 처리	hockey helmet receiving head wearer head crown region left right side regions...
	토큰화	hockey, helmet, receiving, head, wearer, head, crown, region, left, right, side,,...
	단어 사전 구축	'<unk>':0,'<pad>':1,'first':2,'second':3,'one':4,'includes':5,'device':6,'data':7,,..
레이블	원본 데이터	A63B, E06C, B25B
	계층 토큰화	(A,A63,A63B),(E,E06,E06C),(B,B25,B25B)
	단어 사전 구축	'<unk>':0,'<pad>':1,'<sos>':2,'<eos>':3,'G':4,'H':5,'G06':6,'A':7,'B':8,'H04':9,,..

어들 이외의 단어가 출현할 경우 <UNK> 토큰으로 표시를 했으며 입력 시퀀스의 길이보다 짧은 문서인 경우 <PAD> 토큰으로 패딩을 시켰다. 타겟 레이블의 경우 레이블을 계층 단위로 나눠서 단어 사전을 구축했으며 추가적으로 <SOS> 토큰과 <EOS> 토큰을 통해서 레이블의 시작과 끝을 표시했다. 최종적으로 요약문의 단어 사전은 84,987개의 단어, 레이블의 단어 사전은 769개의 단어를 수집할 수 있었다. 또한 IPC 레이블의 경우 시퀀스적 특징을 위해 Main Category를 제외한 Incidental Category는 빈도수를 기반으로 레이블을 정렬했다(Yang et al, 2018).

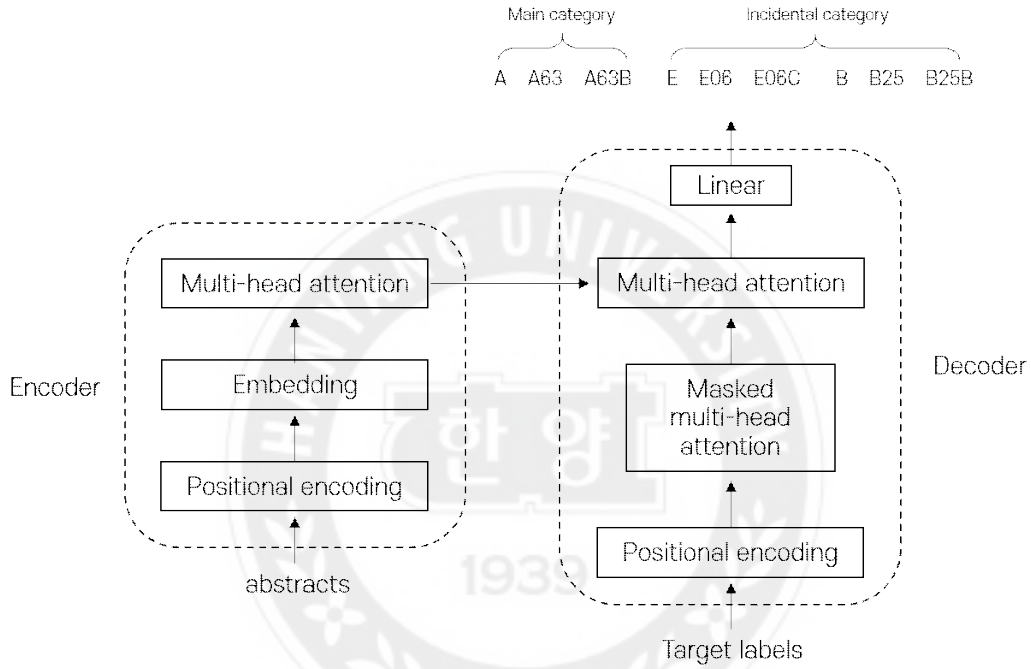
제2절 특허 문서 분류 모델 구현 및 학습

2.1 트랜스포머를 활용한 특허 문서 인코딩

트랜스포머 모델(Transformers)은 2017년 구글에서 발표한 모델로 기존 Sequence to Sequnce 구조인 인코더-디코더를 따르면서 어텐션만으로 구현한 모델이다. RNN 모델을 사용하지 않고 인코더-디코더 구조를 구현했으며 계산 시간 및 성능에 있어서 기존 모델보다 월등하다는 특징이 있다(<그림 4> 참조).

트렌스포머 모델에서 인코더는 셀프 어텐션으로 구성이 되어있다. 셀프 어텐션의 구조는 <그림 5>과 같다. 입력 받은 텍스트 임베딩 값과 서로 다른 가중치 메트릭스 세 개를 통해서 Query, Key, Value 값을 계산되며 계산된 각각의 행렬값은 임베딩된 문서를 세 개의 서로 다른 벡터로 표현했다고 할 수 있다. 셀프 어텐션의 다음 단계는 Query, Key, Value를 통해 각 단어들의 점수를 계산한다. Query와 전치된 Key를 행렬곱을 통해 계산하고 소프트맥스 함수를 통해서 각 단어가 문서에서 어느 정도의 연관성을 가지고 있는지 계산

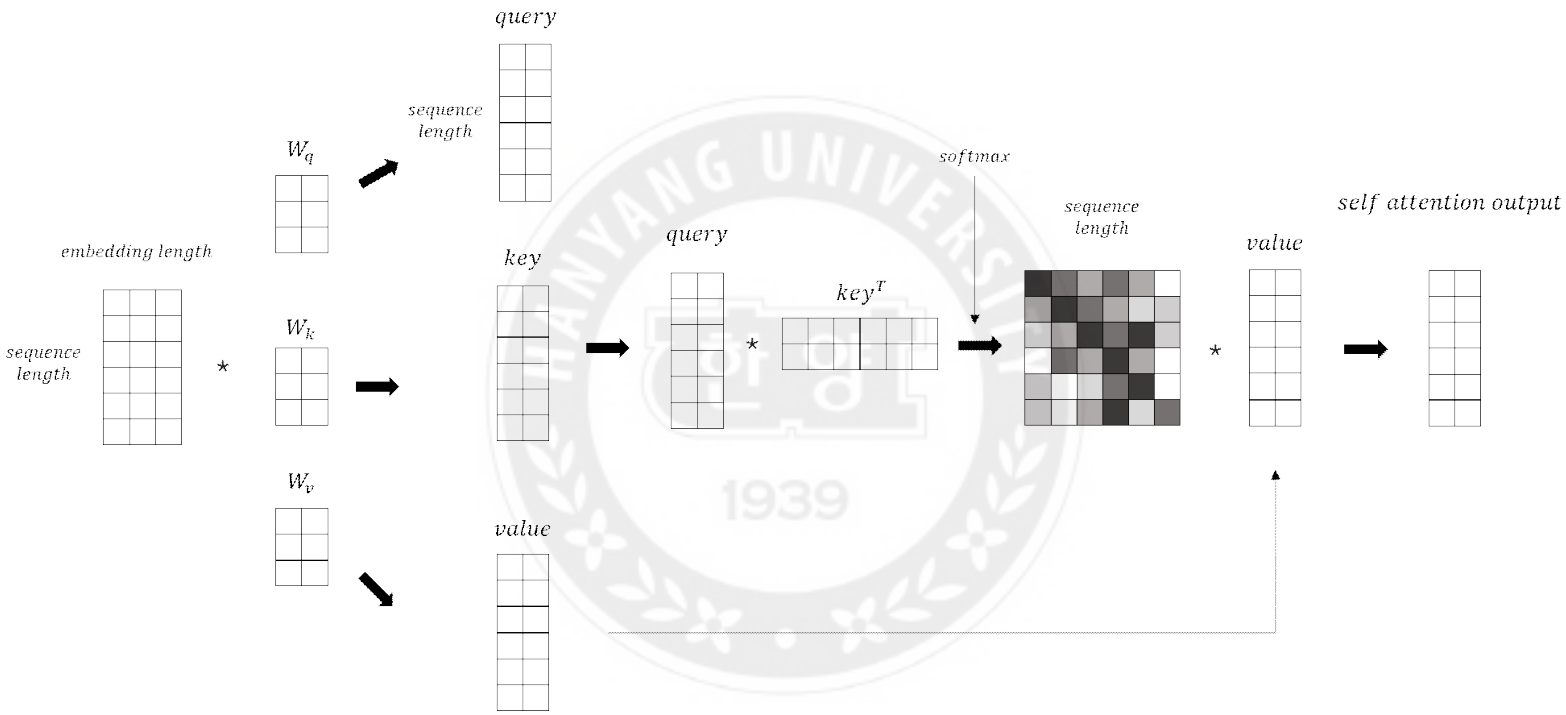
한다(수식 1 참조). 마지막으로 위 같이 계산된 벡터값과 Value를 곱하면 최종적으로 출력된 벡터는 단순이 각각의 단어에 대한 벡터값이 아닌 문서 내에서 단어들이 지닌 의미를 포함한 벡터라고 할 수 있다.



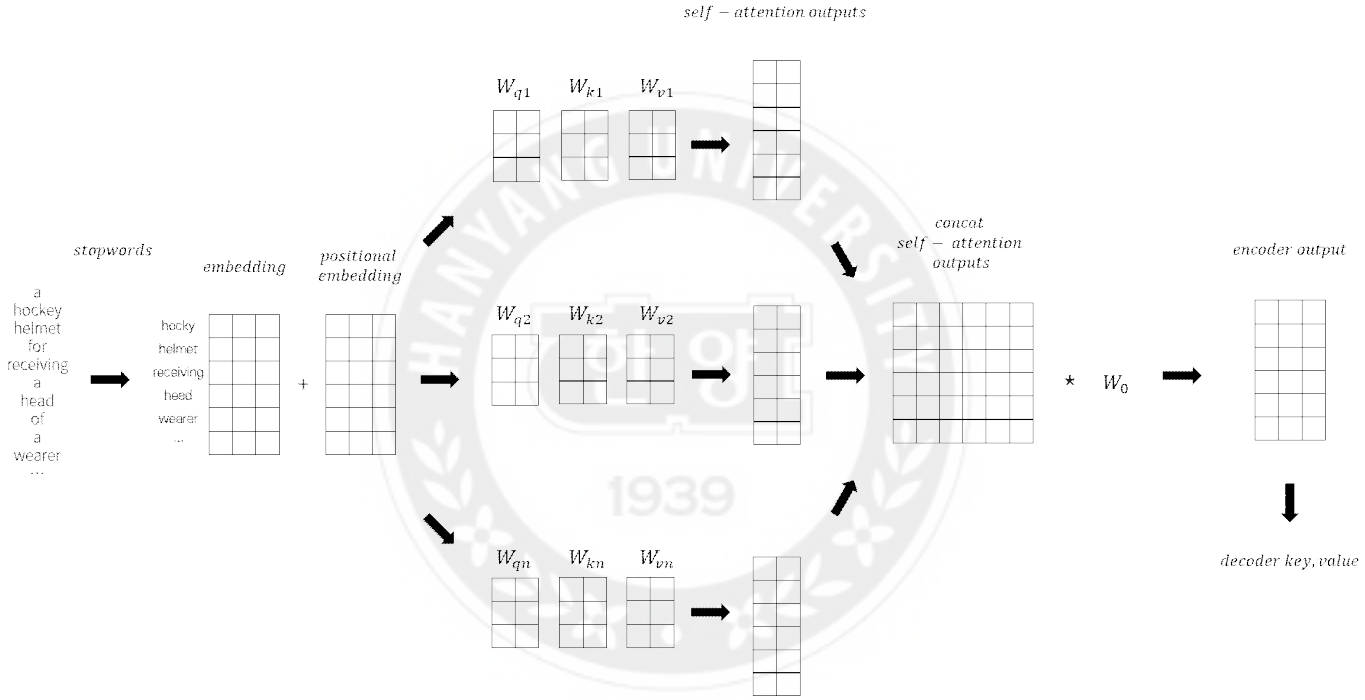
<그림 4> 트랜스포머 모델의 구조

이러한 셀프 어텐션은 본 연구에서 총 8개로 구성하였으며 이를 멀티 헤드 어텐션이라고 부른다(<그림 6> 참조). 멀티 헤드 어텐션에서의 계산은 모두 병렬 처리로 진행되며 마지막에 출력된 셀프 어텐션 값들을 하나로 연결한 후 가중치 메트릭스를 통해 인코더에 입력된 입력값과 동일한 사이즈로 만들어 준다. 마지막 인코더에서 출력된 벡터값은 디코더로 넘어가게 된다.

$$Attention(Q,K,V) = softmax(QK^T/\sqrt{d_k})V \qquad \text{(수식 1)}$$



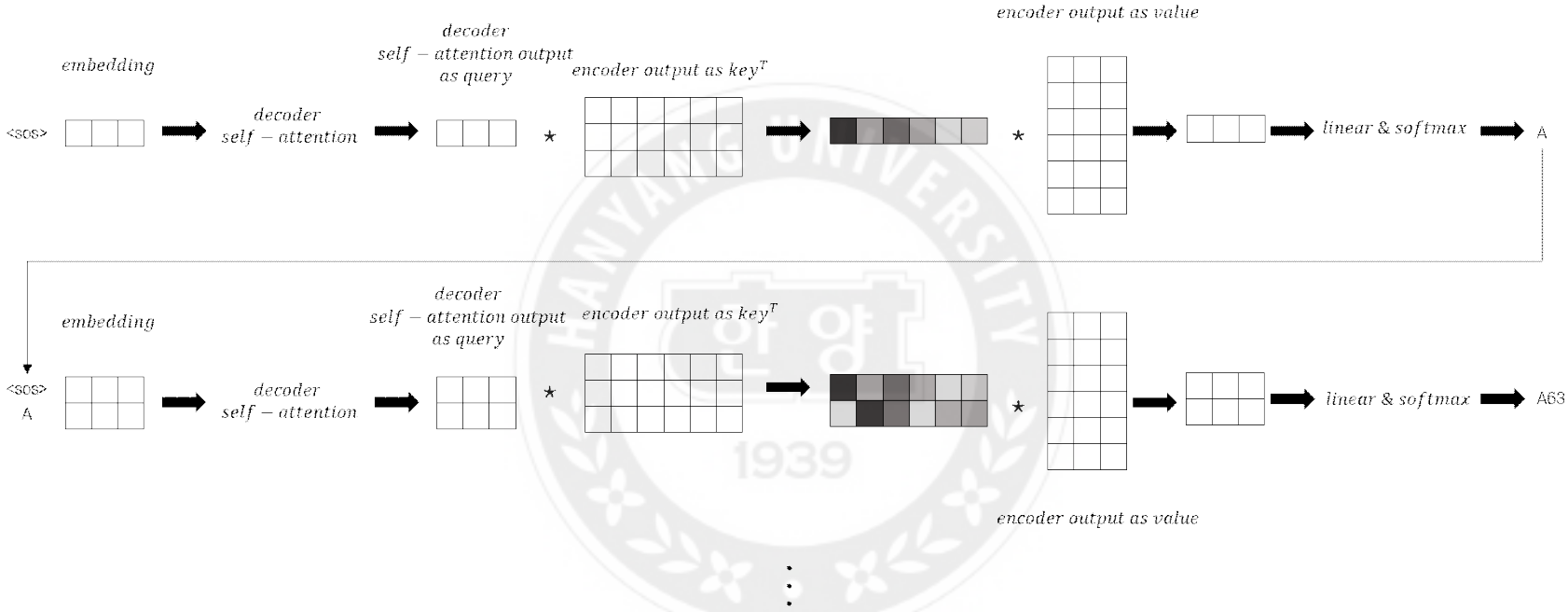
<그림 5> 셀프 어텐션의 구조



<그림 6> 트랜스포머 인코더의 구조

2.2 IPC 분류를 위한 트랜스포머 디코딩

트랜스포머 디코더에서는 인코더와 다르게 마스크드 멀티 헤드 어텐션과 멀티 헤드 어텐션으로 구성되어 있다(<그림 7> 참조). 디코더의 입력값은 타겟 시퀀스의 시작을 알리는 <SOS> 토큰과 레이블 시퀀스이며 인코더에서의 출력값도 함께 디코더에서 사용된다. 우선, 디코더의 첫 입력값으로 임베딩된 <SOS> 토큰과 뒤에 레이블 값이 마스킹된 벡터값이 입력되면 셀프 어텐션을 통해 디코더의 Query 값을 출력한다. 출력된 디코더의 Query는 인코더에서 출력된 값을 Key와 Value로 사용하여 멀티 헤드 어텐션 과정을 거치게 된다. 최종적으로 출력된 벡터값은 선형 함수와 소프트맥스 함수를 거쳐 레이블을 출력하게 된다. 출력된 레이블은 다음 디코더의 입력값으로 들어가게 되고 위와 같은 과정을 반복한다. 하나의 IPC 레이블이 섹션과 클래스, 서브클래스 순서로 출력된 후 다음 레이블을 출력하기 위해 이전 레이블의 서브클래스가 디코더의 입력으로 들어오기 때문에 레이블 간의 의존성을 반영했다고 볼 수 있다. 최종적으로 모델이 출력하는 IPC 레이블은 총 3개이며 이 레이블은 계층적으로 출력하게 된다.



<그림 7> 트랜스포머 디코더의 구조

4. 실험

제1절 평가 데이터셋

모델을 평가하기 위한 데이터셋으로 USPTO-2M 데이터셋을 활용하였다. USPTO-2M 데이터셋은 2006년부터 2015년까지의 미국 특허 데이터를 수집한 것으로 총 2,000,147개의 데이터가 있으며 특허 문서의 제목(Title), 요약(Abstract), 특허번호(No), IPC 레이블(Subclass labels)로 구성되어 있다(<표 4> 참조). 본 연구에서 사용한 데이터는 요약과 IPC 레이블이며 요약 문서는 최대 100 단어까지를 기준으로 전처리를 진행하였으며 단어 사전을 구축할 때 최소 10번 이상 출현한 단어를 기준으로 수집하였다. 결과적으로 요약 문서를 통해서 총 84,987개의 단어를 수집할 수 있었고 IPC 레이블 또한 전처리 과정을 통해서 총 769개의 단어를 수집하였다.

<표 4> USPTO-2M 데이터 예시

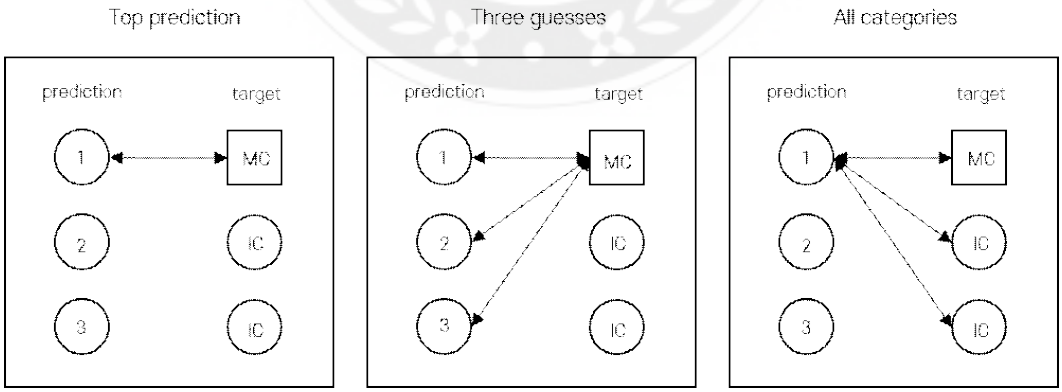
컬럼	데이터
Title	accessory for shoelaces hat brims and the like
Abstract	a decorative and or promotional accessory to be secured to a lace such as a shoelace...
Subclass labels	'A43B', 'A41D', 'A43C'
Number	US08925116

제2절 평가 척도

모델 평가 방법으로는 Fall et al(2003)이 제안한 특허 분류 평가 지표를 사용하였다(<그림 8> 참조). 특허 분류 평가 지표는 micro-average precision을 사용하며 크게 세 가지로 나눠져 있다(수식 2 참조). 첫 평가 지표로는 Top Prediction(TP)이며 모델이 출력한 첫 레이블과 타겟 레이블의 main category와 비교한다. 이 방식은 정확도와 동일하지만 본 연구에서 사용하는 데이터

$$Micro-average\ \ precision = \frac{TruePositive_1 + ... + TruePositive_n}{TruePositive_1 + FalsePositive_1 + ... + TruePositve_n + FalsePositive_n} \quad \text{(수식 2)}$$

특성 상 특정 레이블에 편향됐기에 이 지표 하나만으로 모델을 평가하기에는 다소 무리가 있다. 두 번째 평가 지표로는 Three Guesses(TG)이며 모델이 예측한 세 가지 레이블과 타겟 레이블의 main category와 비교한다. 예측된 세



<그림 8> 특허 문서 분류 평가 지표

가지 레이블 중에서 하나라도 main category와 일치한다면 정답으로 간주한다. 마지막 평가 지표로는 All Categories(AC)로 모델이 예측한 첫 번째 레이블과 타겟 레이블 세 개와 비교하는 방법이다. 예측된 레이블이 세 개의 타겟 레이블 중 하나라도 일치한다면 정답으로 간주한다. 이러한 평가 지표를 사용하는 이유는 특허 데이터 분류가 멀티 레이블 문제이지만 항상 main category 하나를 가지고 있기 때문이다.

제3절 실험 결과

USPTO-2M 데이터는 <표 5>와 같이 연도 별로 데이터를 나눴으며 학습 데이터로는 2006년부터 2013년까지의 데이터, 검증 데이터로는 2014년 데이터, 평가 데이터로는 2015년 데이터를 사용했으며 데이터 수는 각각 1,646,913건, 303,334건, 49,900건이다.

<표 5> 학습, 검증, 평가 데이터 분할

데이터	연도	데이터 수
학습 데이터	2006년~2013년	1,646,913
검증 데이터	2014년	303,334
평가 데이터	2015년	49,900

모델 실험 결과 <표 6>와 같은 결과를 볼 수 있었다. 기존 선행 연구와 비교했을 때 트랜스포머를 활용한 모델이 Top Prediction과 All Categories에서

각각 13.7%, 2.4% 향상된 것을 볼 수 있었고 Three Guesses 지표는 Naive Bayes와 CNN 보다는 우수한 성능을 보였지만 GRU 모델에 비해 6.9% 낮은 것을 확인할 수 있었다.

Three Guesses에서 다소 낮은 성능을 보인 것은 다음과 같이 설명할 수 있다. 모델의 출력값으로 총 3개의 IPC 레이블을 출력하도록 학습이 되었으며 하나의 Main Category와 두 개의 Incidental Category를 출력한다. Three Guesses의 평가 척도는 모델이 예측한 세 개의 레이블 중 하나라도 타겟 레이블인 Main Category에 해당하면 정답으로 간주하는 평가 방법으로 Top Prediction 보다 평가 방법이 유연하여 높은 성능을 보인다. 하지만 모델이 처음으로 출력하는 Main Category 이후의 레이블은 Incidental Category에 해당하며 모델이 예측한 Incidental Category와 타겟 레이블을 비교하기 때문에 성능이 다소 떨어진 것으로 예상된다.

<표 6> USPTO-2M를 활용한 특허 분류 평가 결과

Model	Top Prediction	Three Guesses	All Categories
Naive Bayes	40.1	56.2	53.3
CNN	45.5	67.0	63.4
GRU	54.0	**77.3**	72.7
Transformer with hierarchical label	**67.7**	70.4	**74.3**

위와 같은 결과를 통해서 본 연구는 다음과 같은 학술적, 실무적 의의를 가

진다. 우선, 시퀀스 생성 모델 또한 계층적 멀티 레이블 분류 모델에 적용할 수 있다는 점과 멀티 레이블을 분류하는데 있어서 레이블 간의 의존성을 적용했다는 것에서 학술적 의의를 가진다. 또한. 현실 데이터를 활용하여 연구를 진행했으며 향후 특허 데이터를 분류하는데 있어서 특허 심사에 소요되는 시간 및 비용을 줄일 수 있을 것으로 기대된다.

5. 결론

제1절 결론

기존 특허 문서를 분류하는데 있어서 멀티 레이블 분류는 IPC 레이블의 계층적인 구조를 반영하지 못 했으며 레이블 간의 의존성 또한 적용하지 못한 한계점이 있었다.

본 연구에서는 시퀀스 생성 모델인 트랜스포머를 활용하여 특허 문서를 분류하는데 있어 위와 같은 한계점을 개선하고자 했다. 특허 문서에서 불용어 처리를 진행한 후 IPC 레이블을 계층적으로 전처리를 했으며 단순히 단어 임베딩에서 끝나는 것이 아니라 셀프 어텐션을 통해 문서 내 단어들의 의미를 벡터로 충분히 표현했다. 또한 디코더의 마스크드 셀프 어텐션을 특허 문서를 계층적 레이블로 분류하는 연구를 진행했으며 제시한 방법론이 기존 연구보다 Top Prediction과 All Categories에서 우수한 성능을 보였다는 것을 확인했다.

본 연구에서는 제시한 방법론은 다음과 같은 의의를 갖는다. 첫 번째로 기계번역을 중심으로 활용되는 시퀀스 생성 모델을 계층적 멀티 레이블 분류에 적용하였다. 두 번째로 원-핫 벡터로 분류한 멀티 레이블 모델과는 다르게 시퀀스 생성 모델을 활용하여 레이블 간의 종속성을 포함할 수 있었다. 마지막으로 모델 학습을 위해 정비된 데이터가 아닌 대량의 현실 데이터를 활용하여 모델을 구현했으며 특허 분류를 진행하는데 있어서 시간 및 비용을 줄여줄 것으로 기대된다.

제2절 향후 연구

향후 연구로는 IPC 레이블은 본 연구에서 다룬 서브클래스 레벨보다 더 세분화된 그룹과 서브그룹까지 구성이 되어있기 때문에 이를 적용하기 위해서는 더 세분화된 단위의 IPC 레이블에 대해 모델링할 필요가 있다. 또한 최근 자연어 처리 분야에서 텍스트 임베딩 방법으로 BERT (Bidirectional Encoder Representations from Transformers)가 다양한 자연어 처리 테스크에서 좋은 성능을 보이고 있기 때문에 이를 활용할 수 있는 방안을 모색할 필요가 있다. 마지막으로 어텐션 메커니즘을 활용하여 문서의 키워드를 추출할 수 있다면 단순히 분류에서 그치지 않고 문서 분류 결과에 대해 해석을 제공할 수 있을 것으로 생각된다.

참 고 문 헌

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." arXiv preprint arXiv:1409.0473 (2014).

Chen, Yangchi, Melba M. Crawford, and Joydeep Ghosh. "Integrating support vector machines in a hierarchical output space decomposition framework." In IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium, vol. 2, pp. 949-952. IEEE, 2004.

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." arXiv preprint arXiv:1406.1078 (2014).

Choi, Seokkyu, Hyeonju Lee, Eunjeong Lucy Park, and Sungchul Choi. "Deep Patent Landscaping Model Using Transformer and Graph Embedding." arXiv preprint arXiv:1903.05823 (2019).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

Fall, Caspar J., Atilla Törcsvári, Karim Benzineb, and Gabor Karetka. "Automated categorization in the international patent classification." In Acm Sigir Forum, vol. 37, no. 1, pp. 10-25. New York, NY, USA: ACM, 2003.

Grawe, Mattyws F., Claudia A. Martins, and Andreia G. Bonfante.

"Automated patent classification using word embedding." In 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 408-411. IEEE, 2017.

Guyot, Jacques, Karim Benzineb, Gilles Falquet, and Simple Shift. "myClass: A Mature Tool for Patent Classification." In CLEF (notebook papers/LABs/workshops). 2010.

Kim, Yoon. "Convolutional neural networks for sentence classification." arXiv preprint arXiv:1408.5882 (2014).

Larkey, Leah S. "A patent search and classification system." In Proceedings of the fourth ACM conference on Digital libraries, pp. 179-187. 1999.

Li, Shaobo, Jie Hu, Yuxin Cui, and Jianjun Hu. "DeepPatent: patent classification with convolutional neural networks and word embedding." Scientometrics 117, no. 2 (2018): 721-744.

Li, Zhen, Derrick Tate, Christopher Lane, and Christopher Adams. "A framework for automatic TRIZ level of invention estimation of patents using natural language processing, knowledge-transfer and patent citation metrics." Computer-aided design 44, no. 10 (2012): 987-1010.

Liu, Jingzhou, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. "Deep learning for extreme multi-label text classification." In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 115-124. 2017.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. "Efficient estimation of word representations in vector space." arXiv preprint

arXiv:1301.3781 (2013).

Minaee, Shervin, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. "Deep learning based text classification: A comprehensive review." arXiv preprint arXiv:2004.03705 (2020).

Peng, Hao, Jianxin Li, Yu He, Yaopeng Liu, Mengjiao Bao, Lihong Wang, Yangqiu Song, and Qiang Yang. "Large-scale hierarchical text classification with recursively regularized deep graph-cnn." In Proceedings of the 2018 World Wide Web Conference, pp. 1063-1072. 2018.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532-1543. 2014.

Risch, Julian, and Ralf Krestel. "Learning Patent Speak: Investigating Domain-Specific Word Embeddings." In 2018 Thirteenth International Conference on Digital Information Management (ICDIM), pp. 63-68. IEEE, 2018.

Roudsari, Arousha Haghighian, Jafar Afshar, Charles Cheolgi Lee, and Wookey Lee. "Multi-label Patent Classification using Attention-Aware Deep Learning Model." In 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 558-559. IEEE, 2020.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin.

"Attention is all you need." Advances in neural information processing systems 30 (2017): 5998-6008.

Wu, Chih-Hung, Yun Ken, and Tao Huang. "Patent classification system using a new hybrid genetic algorithm support vector machine." Applied Soft Computing 10, no. 4 (2010): 1164-1177.

Xue, Gui-Rong, Dikan Xing, Qiang Yang, and Yong Yu. "Deep classification in large-scale text hierarchies." In Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 619-626. 2008.

Yan, Yan. "Hierarchical Classification with Convolutional Neural Networks for Biomedical Literature." International Journal of Computer Science and Software Engineering 5, no. 4 (2016): 58.

Yang, Pengcheng, Xu Sun, Wei Li, Shuming Ma, Wei Wu, and Houfeng Wang. "SGM: sequence generation model for multi-label classification." arXiv preprint arXiv:1806.04822 (2018).

Yang, Zichao, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. "Hierarchical attention networks for document classification." In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pp. 1480-1489. 2016.

Abstract

Hierarchical multi-label classification of patent documents using sequence generation model

Yoon, Seung Joo
Dept. of Business Informatics
The Graduate School of
Hanyang University

Many real-world documents can be classified in hierarchical structure. For example, an electronic documents can be divided into different categories and each category can be subdivided into categories of multiple lower layers. In particular, for the patent data, the number of labels that can classify documents based on the lowest level is very large, and various approaches are under way to achieve more accurate classification.

Previous researches have approached patent classification as a multi-label classification problem. However, these methods ignore dependencies between labels and do not fully apply the hierarchical structure of the International Patent Classification (IPC) label.

To overcome these limitations, this research introduce a hierarchical multi-label classification model using Transformers, which has shown good performance in machine translation. Using the patent's summary documentation, IPC labels were classified into sections, classes, and subclasses, and could include relationships between multiple labels that were previously lacking in patent classification studies.

This study used USPTO-2M data to evaluate performance on patent classification and ultimately found that the classification performance in 'Top Prediction' and 'All Categories' is superior to existing models.

This research

Keyword Patent classification, generative model, machine translation, hierarchical classification, multi-label classification

연구 윤리 서약서

본인은 한양대학교 대학원생으로서 이 학위논문 작성 과정에서 다음과 같이 연구 윤리의 기본 원칙을 준수하였음을 서약합니다.

첫째, 지도교수의 지도를 받아 정직하고 엄정한 연구를 수행하여 학위논문을 작성한다.

둘째, 논문 작성시 위조, 변조, 표절 등 학문적 진실성을 훼손하는 어떤 연구 부정행위도 하지 않는다.

셋째, 논문 작성시 논문유사도 검증시스템 "카피킬러"등을 거쳐야 한다.

2020년12월21일

학위명 : 석사

학과 : 비즈니스인포매틱스학과

지도교수 : 김종우

성명 : 윤승주 (서명)

한 양 대 학 교 대 학 원 장 귀 하

Declaration of Ethical Conduct in Research

I, as a graduate student of Hanyang University, hereby declare that I have abided by the following Code of Research Ethics while writing this dissertation thesis, during my degree program.

"First, I have strived to be honest in my conduct, to produce valid and reliable research conforming with the guidance of my thesis supervisor, and I affirm that my thesis contains honest, fair and reasonable conclusions based on my own careful research under the guidance of my thesis supervisor.

Second, I have not committed any acts that may discredit or damage the credibility of my research. These include, but are not limited to : falsification, distortion of research findings or plagiarism.

Third, I need to go through with Copykiller Program(Internet-based Plagiarism-prevention service) before submitting a thesis."

DECEMBER 21, 2020

Degree : Master

Department : DEPARTMENT OF BUSINESS INFORMATICS

Thesis Supervisor : Jong Woo Kim

Name : YOON SEUNGJOO (Signature)