계층적 주의 네트워크를 활용한 특허 문서 분류

장현철*, 한동희*, 류태선*, 장형국* 임희석**
*고려대학교 컴퓨터정보통신대학원 빅데이터융합학과
**고려대학교 컴퓨터학과
e-mail : justinjang87@korea.ac.kr, ehdgml76@korea.ac.kr, yts1993@korea.ac.kr,
gagagiga@korea.ac.kr, limhseok@korea.ac.kr

Patent Document Classification by Using Hierarchical Attention Network

Hyuncheol, Jang*, Donghee, Han*, Teaseon Ryu*, Hyungkuk Jang*, HeuiSeok Lim**

*Dept. of Big Data Convergence, Korea University Graduate School of Computer and Information Technology
**Dept. of Computer Science and Engineering, Korea University

요 약

최근 지식경영에 있어 특허를 통한 지식재산권 확보는 기업 운영에 큰 영향을 주는 요소이다. 성공적인 특허 확보를 위해서, 먼저 변화하는 특허 분류 체계를 이해하고, 방대한 특허 정보 데이터를 빠르고 신속하게 특허 분류 체계에 따라 분류화 시킬 필요가 있다. 본 연구에서는 머신 러닝 기술 중에서도 계층적 주의 네트워크를 활용하여 특허 자료의 초록을 학습시켜 분류를 할 수 있는 방법을 제안한다. 그리고 본 연구에서는 제안된 계층적 주의 네트워크의 성능을 검증하기 위해 수정된 입력데이터와 다른 워드 임베딩을 활용하여 진행하였다. 이를 통하여 특허 문서 분류에 활용하려는 계층적 주의 네트워크의 성능과 특허 문서 분류 활용화 방안을 보여주고자 한다. 본 연구의 결과는 많은 기업 지식경영에서 실용적으로 활용할 수 있도록 지식경영 연구자, 기업의 관리자 및 실무자에게 유용한 특허분류 기법에 관한 이론적 실무적 활용 방안을 제시한다.

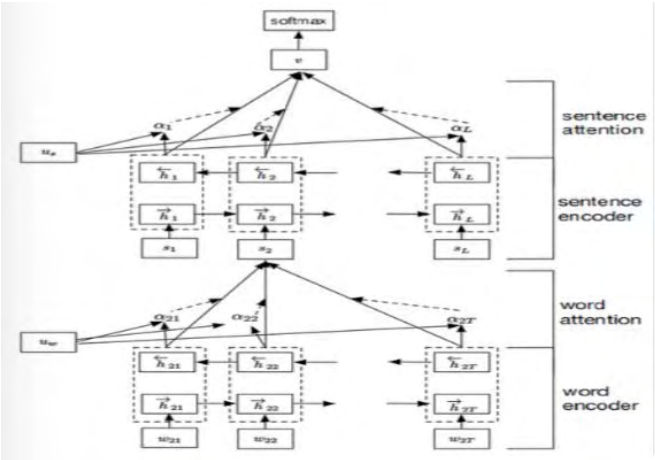
1. 서론

Document Classification이라 불리는 문서 분류는 임의의 텍스트를 정해진 범주에 따라 분류하는 것을 일컫는다. 인터넷의 발전과 정보 기술의 발전에 따라 전산화된 문서의 양이 점점 증가하고 있고, 이에 따른 정보의 분류 문제 역시 중요한 문제로 제기되고 있다. 문서 중 일부분의 문장, 문단, 단어가 경제 분야인지, 정보통신 분야인지 그리고 정보통신 분야 중에서는 통신기술 분야인지 반도체 분야 인지에 대한 분류를 손쉽고 정확하게 수행하고자 하는 요구사항이 많이 생겨났다. 데이터 마이닝을 통해 의미 있는 키워드, 패턴, 지식을 분석하여 보다 유용한 정보를 추출할 수 있다. 이러한 목적으로 지난 몇 년간 각종 논문에서 사용한 CNN, RNN, 계층적 주의 네트워크, 하이브리드 네트워크 등 여러가지의 머신러닝(기계학습) 기술을 활용한 연구가 활발하게 이루어지고 있으며 본 연구에서는 이중 계층적 주의 네트워크를 연구하고자 한다. 그리고 제2장에서는 실제 본 연구에서 활용하고자 하는 연구 모델 소개를 간략하게 수행하였다. 제3, 4장에서는 연구 모델의 성능 향상 방안으로 워드 임베딩 방식 중에 fastText, GloVe라는 모델을 설명하고, 인풋 데이터를 가공하며 실험 결과를 제시한다. 그 다음으로 제5, 6장에서는 실질적으로 본 연구의 목적인 특허 문서의 분류를 위해 계층적 주의 네트워크를 활용하여 머신러닝(기계학습)을 수행하고, 결과를 제시한다. 마지막으로 7장에서는 결론과 앞으로의 향후 연구 방향에 관해 서술한다. 이러한 문서분류 머신러닝(기계학습) 연구의 끝에는 결과 데이터를 통해 앞으로 기업 내에서 활용될 수 있는 특허 분류 체계 구축의 초석을 마련하고자 한다.

2. 활용하고자 하는 연구모델 소개

특허 문서 분류에 활용한 기계학습 기법인 문서 분류를 위한 계층적 주의 네트워크의 전체 구조는 그림 1과 같으며 단어 시퀀스 인코더, 단어 단계의 주의 레이어, 문장 인코더 및 문장 주의 계층으로 구성이 된다. Hierarchical Attention Network 모델(이하

HAN 모델)[1]은 문서 수준의 분류에 중점을 두고 모델이다. 각 문서에 일반적인 A라는 문장이 있다고 가정한다면 각 문장은 단어들로 구성이 되어 있을 것이다. 각 문장을 또한 각 단어 벡터로 투영한다. 이러한 벡터 표현에서 문서 분류를 수행하는 분류기준을 구성하고 다음에서는 계층적 구조를 사용하여 워드 벡터로부터 문서 레벨 벡터를 점진적으로 작성하는 방법을 HAN 모델은 소개해주고 있다.[1]



(그림 1) hierarchical Attention network 모델 구조

3. 모델 성능 향상 방안 1 (워드임베딩)

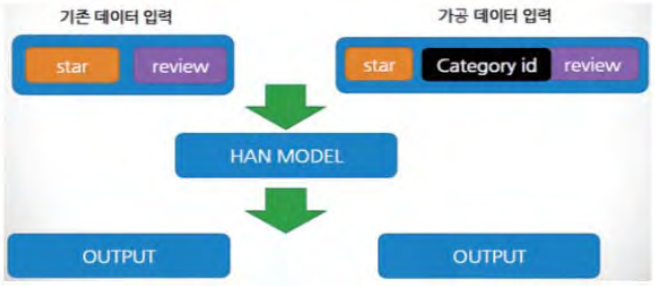
연구 모델인 HAN의 성능 향상을 위해 기존의 워드임베딩 방식은 word 2 vec를 활용하여 임베딩하는 방식이지만 본 연구에서는 pre-trained word vector를 사용하였다. 2가지 모델을 활용하였는데 먼저 Twitter 데이터를 활용한 Glove 그리고 Wikipedia 데이터를 활용한 fastText를 활용하였다. 간략히 두 가지 모델을 설명하면 아래 표와 같다.

GloVe[2]	fastText[3]
• 스탠포드 컴퓨터 과학과에서 제작. • 글로벌 행렬 분해 모델과 local context window 모델을 결합하여 Wikipedia, twitter 단어 벡터를 학습한다.	• Word2Vec 을 직접 제작 설계한 Mikolov 가 최근 2016 년에 제안한 Embedding 방식이다. • 원래 기존 단어를 부분 단어(subword)의 벡터들로 표현이다.

이처럼 미리 훈련된 워드 벡터를 활용하여 현재 연구 모델의 정확도 향상을 도모하였다.

4. 연구 모델 성능 향상 방안 2(인풋데이터 가공)

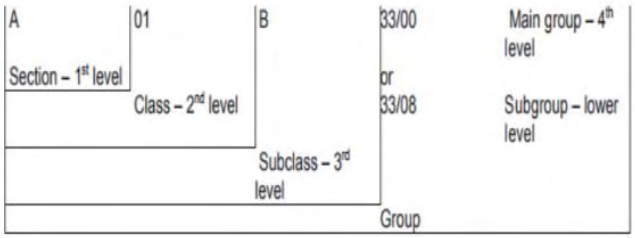
다음으로 진행한 방식은 입력 데이터에 가공을 진행하였다. 연구 모델인 계층적 주의 네트워크의 정확도 향상을 위해서 입력데이터의 방식에서 특정 키워드를 추가하는 방식이다. 이 방식은 현재 계층적 주의 네트워크 내에서 키워드와 실제 본문 내용의 상관관계가 이루고 그를 통하여 실제 별점 추측에 상관관계가 발생할 것으로 예측되어 진행한 방안이다. 자세한 내용은 아래 그림을 보면 이해하기가 쉽다.
아래 그림을 보면 기존 입력 데이터의 형식은 본문의 내용(Review)과 별점(Star) 두 개의 구성 요소로 이루어져 있다. 하지만 본 연구에서 성능 향상을 위해 진행한 방식은 내용(Review)과 별점(Star) 그리고 특정 키워드 (Category Id)를 추가하였다.



(그림 2) 데이터 가공 방식 구조

5. 연구 모델을 특허 분류에 적용

본 연구 모델을 활용하여 특허 문서 분류를 하고자 한다. 특허 문서 분류 체계는 IPC1, CPC, F-term, USPC, ECLA 등 분류 목적이나 활용에 따라 다양하게 사용하고 있으며, 분류 체계의 기본적인 존재 이유는 효과적인 검색 도구 (World Intellectual Property Organization, 2017)로 활용하기 위함이다. 분류 코드 배열은 Section, Class, Subclass, Main Group, Sub Group으로 아래와 같이 계층적인 구조를 가진다.



(그림 3) COMPLETE CLASSIFICATION SYMBOL

최근에는 IPC[4]의 단점을 개선하고자 EPO와 USPTO가 협력하여 새로운 분류 체계인 CPC[5]가 도입되었으며, 검색 도구로서의 효과를 극대화하기 위해 전체 상징이 26만 개로 세분화(IPC는 7만개)

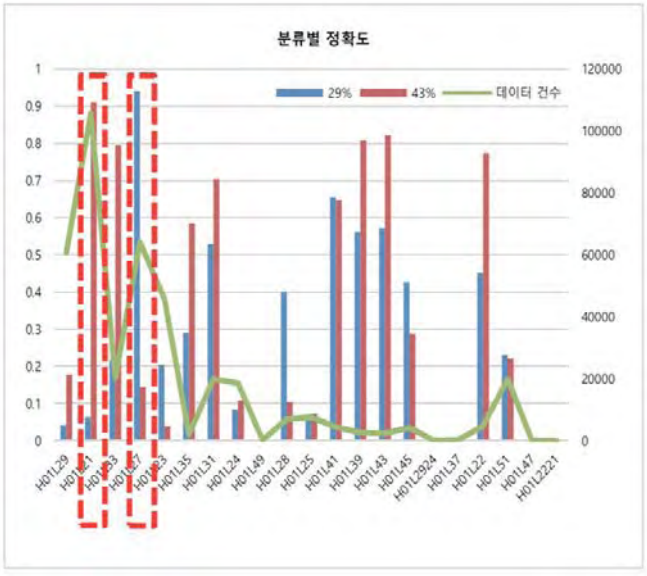
되었다. CPC 코드는 IPC 표준과 동일한 형태의 Main Trunk와 2000시리즈, Y-부문으로 구분된다. CPC 분류 역시 IPC 체계와 마찬가지로 계층화 (EPO and USPTO, Version 1.0 Revision 4.00)되어 있으며, 특허 문서 분류를 한 번에 26만 개 중 하나로 분류할 것인지 혹은 부문, 클래스 등으로 세분되는 계층적 구조를 따라 모델을 설계할 것인지에 대한 의사 결정이 필요하다. 한 사람이 26만 개에 달하는 상징을 모두 숙지하고 분류하는 것이 현실적으로 어려운 일이므로, 담당 기술에 따라 여러 분류 담당자가 일정 기술 범위를 전담하여 최종 분류를 수행하고 있으며 기계 학습 기반의 분류 모델도 유사한 구조로 설계할 수 있다. 1단계로 부문 분류 모델을 구축하고, 2단계로 동일한 모델을 사용하되 부문마다 각 부문으로 분류된 데이터만으로 학습시킨 부문별 클래스 분류 모델을 만드는 방식으로 계층화된 모델 학습이 가능하다. 다만 하위 레벨로 내려갈수록 사용 가능한 데이터 감소 및 그에 따른 정확도 저하가 일어날 수 있다. 본 연구에서는 계층적 주의 네트워크를 활용하여 Main Group 레벨의 분류를 시도하고자 한다. Main Group 레벨의 분류를 자동화한다면, 담당자 업무분장 등 부분적인 분류 프로세스 자동화에 활용함으로써 효율화를 도모할 수 있다. 본 연구는 USPTO에서 무상으로 공급하는 공보 데이터 중 제1 분류가 반도체 관련인 특허(Subclass H01L) 중 요약 텍스트가 존재하는 486,355건을 대상으로 한다. 주의 계층에서 분류에 영향을 얼마나 끼치는지에 따라 단어나 문장의 가중치가 영향을 받을 수 있을 것이므로 스톱워드 제거 등의 전처리는 수행하지 않았다.

6. 연구 결과

위의 3 가지 연구법에 따라 진행한 연구 결과는 아래의 표와 같으며 목적 분류는 다음과 같이 분류하였으므로 참조 바란다.
(1-코드분석, 2-데이터가공, 3-학습시간단축, 4-평점분류, 5-특허분류)

목적	데이터	데이터 수	파라미터	훈련 정확도	개발 정확도
1,2	YELP2013	468,260	epoch 3, lr=0.001	0.73	0.54
1	YELP2013	468,260	epoch 3, lr=0.001		0.55
1	YELP2013	468,260	epoch 3, lr=0.001		0.63
3	YELP2013 (학습시간측정)	97,621	epoch 3, lr=0.001	0.64	0.39
3	YELP2013 (학습시간측정)	97,621	epoch 5, lr=0.001	0.74	0.57
4	YELP2013 (GloVe)	218,656	epoch 3, lr=0.001	0.47	0.45
5	YELP2013 (데이터가공)	500	epoch 3, lr=0.001	0.30	0.33
5	YELP2013 (fastText)	218,656	epoch 9, lr=0.001	0.83	0.37
5	USPTO(GloVe)	486,355	epoch 3, lr=0.001	0.44	-
5	USPTO	486,355	epoch 3, lr=0.001	0.74	0.29
5	USPTO	486,355	epoch 3, lr=0.001	0.70	0.43
5	USPTO	486,355	epoch 6, lr=0.001	0.73	0.62

본 연구 결과를 정리하자면 첫째로 감성 분류는 평점 1과 5 외에는 분류 성능저하가 나타난다. Class 별마다 학습 데이터 수와 정확도가 비례하는 현상은 발생하지 않았으며, 최저점과 최고점에 대한 예측 정확도는 높았으나 중간 점수 예측은 상대적으로는 정확도가 낮다. 둘째로 특허 분류 정확도는 62%을 기록했으나, 심볼마다 정확도가 실험 마다 큰 편차를 가진다.



(그림 4) **특허 분류 실험에서 심볼 정확도 편차 발생**

심볼에 따른 편차가 발생하는 현상은 반도체 분류에서 대표적인 Main Group 인 H01L21(반도체 제조방법), H01L27(기판상에 제조된 부품 혹은 장치)의 경우를 보면 알 수 있다. 29%을 기록한 실험에서 H01L21 분류에 그다지 성공하지 못했지만 H01L27 은 매우 높은 정확도를 보여주고 있다. 반대로 43%를 기록한 실험에서는 오히려 H01L27 의 정확도가 매우 낮은 것을 알 수 있다.

이러한 현상을 제거하기 위해서는 본 특허 문서가 “제조방법이 기재된 물건 발명(PBP)”과 같이 물건 발명의 구조적 구성보다 제조방법이 주된 요소인지에 대해 식별할 수 있어야 하므로, 발명의 형태와 그 제조 방법이 함께 기재되는 요약 텍스트를 사용하여 분류하는 것보다는 청구항을 순서대로 활용하여 실험하는 것이 바람직할 것으로 판단된다. 또한 GloVe 의 Wikipedia 기반 pre-trained 벡터를 사용한 시도는 성공하지 못하였다.

특허 문서로부터 발생 빈도순으로 5 만개의 단어로 단어 사전을 생성하고, 임베딩 벡터를 GloVe 의 벡터로 대체하였으나 19,964 개만 일치할 정도로 반도체 관련 특허와 Wikipedia 의 단어가 일치하는 정도가 높지 않았다. 이는 스톱워드 제거와 같은 전처리를 수행하지 않은 것도 하나의 원인으로 생각된다. 하지만 스톱워드 제거를 사람의 판단에 의존하지 않고 궁극적으로 기계 학습을 통해 해결하기 위해서는 다른 방식의 일치율 개선 방식에 대한 검토가 필요하다.

7. 결론 및 향후 연구

본 연구 결과에서는 계층적 주의 네트워크를 활용하되 입력 데이터 가공과 워드 임베딩 방식의 변화를 통하여 성능 향상을 도모하였다. 그리고 실질적인 활용을 위해서 특허문서 분류를 진행하였다. 결과적으로 입력 데이터 가공을 통한 성능 향상이 기대만큼 이뤄지지는 못하였다. 또한, 워드 임베딩 방식을 기존 Word2Vec 에서 fastText 로 변경한 시도 역시 향상하지는 못하였다.

앞으로 성능을 개선하기 위한 연구가 필요하며, 단어 사전의 일치율 개선을 위하여 3 가지 실험을 진행하고자 한다. 첫째로, Tokenizer 변경 또는 업그레이드, 둘째로 스톱 워드 제거 혹은 품사 한정 등 데이터 전처리, 세 번째로 GloVe 에 수록된 단어만으로 단어 사전을 구축하거나, 일치하는 단어의 벡터는 교체하되 이외의 벡터는 랜덤하게 초기화 후 학습하는 등의 방식을 통해 정확도 개선이 가능한지 실험하고자 한다. 또한, 계층적 주의 네트워크를 활용한 특허 문서 분류는 실험마다 심볼 정확도가 달라지는 문제를 해결하기 위해 입력 데이터를 청구항으로 확장할 필요가 있다.

마지막으로 방대한 특허 문서 전체를 기반으로 Word2Vec 혹은 fastText 등의 방법으로 Pre-Trained Word Vector 를 구축하는 연구를 수행할 필요가 있다.

참고문헌

[1] Zichao Yang1, ”Hierarchical Attention Networks for Document Classification,” HLT-NAACL, 2016

[2] Jeffrey Pennington, ”GloVe : Global Vector for Word Representation,” Computer science Department, Stanford University, Stanford, CA 94305 , 2014

[3] Tomas Milcolov, ”Bag of Tricks for Efficient Text Classification,” Facebook AI Research, 2016

[4] Guide to the International Patent Classification, WIPO, 2017

[5] Guide to the Cooperative Patent Classification, EPO and USPTO, 2017