인공지능과 자연어 처리를 통한 IPC 기술 분류

고려대학교 산업경영공학부 금강산, 김현진, 유선호

지식재산융합인재양성사업단 지식재산캠프

고려대학교 산업경영공학부

Table of contents

1. 연구 배경

연구 배경 및 필요성

➢ **연구 목표**: 본 연구는 특허 빅데이터 분석을 통한 기술평가 연구로, 특허 DB를 기반으로 “기술의 원천성 및 가치 판단”을 목표로 함.

연구 흐름1

분야/코드 식별의 어려움

특허 및 관련 문서는 해당 기술의 정확한 분야 및 IPC/CPC 코드 식별 어려움

연구 필요성

연구 흐름2

인공지능/자연어 처리 적용의 유용성

특허 문서, 기술사업계획서를 딥러닝/머신러닝에 적용해 주요 핵심 키워드 도출 및 IPC/CPC 코드 식별 가능

연구 목적성

연구 흐름3

도출 키워드 및 코드의 활용성

도출된 IPC/CPC 코드와 핵심 키워드 통해 기술 분야 현황 분석, 기술 원천성 파악

연구 활용 가능성

1. 연구 배경

활용 가능 자연어 처리 모델 목록

➢ 아래 5가지 대표적 모델 중 KoBERT 모델 채택

모델명	특징
KoBART Korean Bidirectional and Auto-Regressive Transformers	입력 데이터 일부 노이즈 추가, 원문으로 복구 40 GB 이상 한국어 텍스트 학습
DNN Deep Neural Network	인공신경망 기반 자연어 처리 알고리즘, 입력층과 출력층 사이 여러 개 은닉층 보유
KoBERT Korean Bidirectional Encoder Representations from Transformer	구글 BERT 알고리즘 기반으로 40GB 이상 한국어 텍스트 학습, 대형 인공지능, 오픈소스
Google T5	모든 텍스트 기반 언어 문제 텍스트 대 텍스트 형식으로 변환하는 통합 프레임워크 도입
Kopatbert	특허 분야 특화 고성능 사전학습, 한국어 특화 모델

2. 세부 연구 내용

KoBERT 모델 선택 이유

- SKT Brain KoBERT Model : 한국어 버전 자연어 처리 모델
- BERT : Bidirectional Encoder Representations from Transformer 약자로, 텍스트 양방향(앞뒤)로 확인해 자연어 처리하는 모델
- 한국어의 경우 타 언어보다 복잡도가 높기에 한국어 특화 자연어 처리 모델 사용이 필요
- 한국어에 대해 많은 사전 학습 이루어져 있고 다중 분류 가능한 것이 강점이기 때문에, KoBERT 모델 이용

2. 세부 연구 내용

KoBERT 모델 구현

- 주어진 Dataset 중 필요한 부분을 학습 Dataset과 테스트 Dataset으로 나누어 분류, 비율은 7 : 3으로 설정
- BERT 모델의 기능을 활용해 문장을 토큰별로 쪼개 저장하는 Tokenizer 기능 활용
- Max_len(학습할 데이터의 최대 길이), batch_size(트레이닝 Dataset 나눈 소그룹 데이터수), num_epochs(전체 트레이닝 Dataset 신경망 통과 횟수 = 학습 횟수) 등 Parameter 조정해가며 학습 방식 조정

2. 세부 연구 내용

학습 결과 분석

➢ 데이터 분석 결과 메인 ipc 코드를 기준으로 데이터의 개수가 100개 이상인 상위 4개 항목에 대하여 학습 및 테스트 진행

➢ Max_len = 256, Batch_size = 32, Epoch = 10

```
100% ████████████████████████████████████████ 17/17 [00:23<00:00, 1.21s/it]
epoch 10 batch id 1 loss 0.1405385434627533 train acc 0.96875
epoch 10 train acc 0.9926470588235294
100% ████████████████████████████████████████ 8/8 [00:03<00:00, 2.00it/s]
epoch 10 test acc 0.890625
```

테스트 데이터에 대해
89% 정도의 정확도를 보임

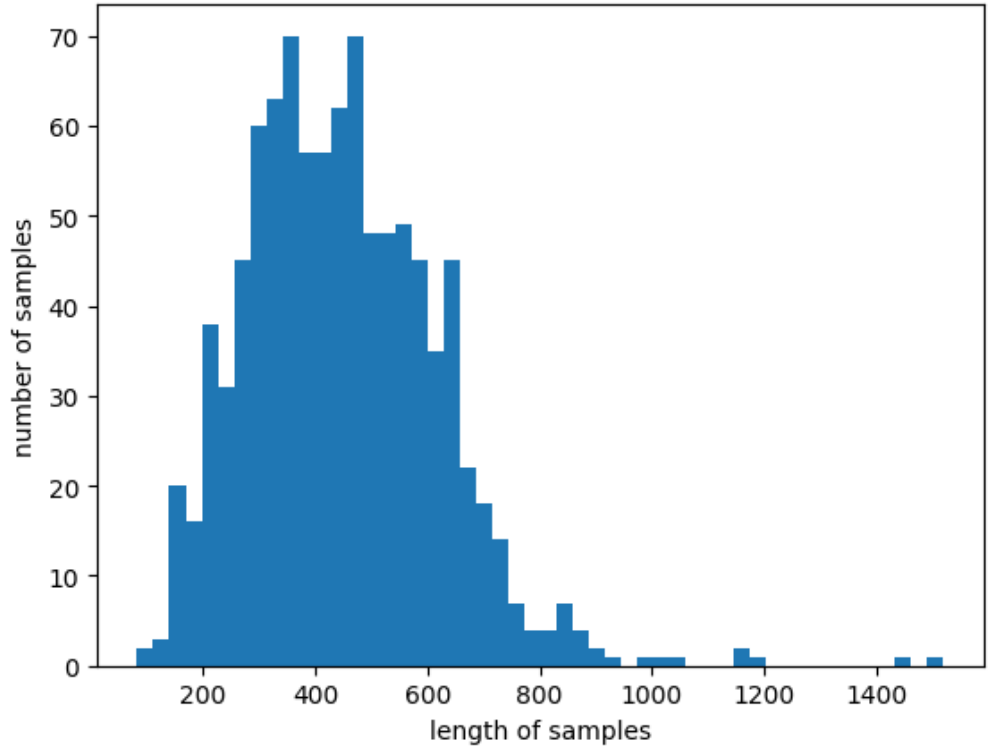
2. 세부 연구 내용

모델 파라미터 분석

- 기존 설정값 : Max_len = 64, Batch_size = 64, Epoch = 10
- 데이터 분석 결과 리뷰의 최대 길이: 1538, 평균 길이 : 477.4
- 설정한 Max_len 값에 의해 데이터의 앞부분만 학습,
  대부분 데이터에 대해 일부 내용만 학습하고 나머지 내용
  누락한다는 문제점 존재

  → Max_len의 증가를 통해 문제 개선 시도 ,

  작동 환경 상 최댓값인 256 이용



2. 세부 연구 내용

원인 분석

➢ 나온 결과에 대한 해석 : 정확도가 높은 이유는 메인 ipc 코드를 데이터 개수 100개 이상인 상위 4개까지만 포함했기 때문

→ 학습 및 테스트에 포함할 메인 ipc 코드 종류 증가시켜야 더욱 실용적인 모델 구현 가능

➢ 데이터 자체의 분포 방식 고려 필요

→ 데이터 특성에 대한 분석 진행 후 보정 과정 진행

2. 세부 연구 내용

데이터 특성 분석

- 메인 ipc 코드별 데이터 개수 상위 10개 목록
- 데이터가 가장 많은 G07F와, 10번째로 많은 G09F는 232개의 차이를 보임
- 전체 1,286개 데이터 중 G07F는 20.4%, G09F는 2.3%를 차지하며

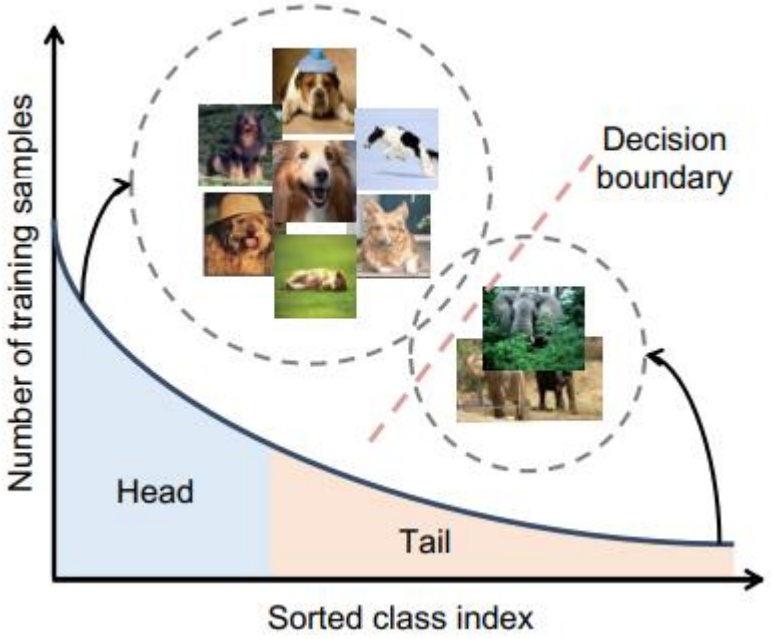
  데이터의 분포가 고르지 않음을 알 수 있음

메인 ipc 코드	개수
G07F	262
G07D	196
G01G	181
G06Q	119
G05D	72
G16C	69
G05B	56
G07C	38
G08G	34
G09F	30

2. 세부 연구 내용

데이터 특성 분석

- Long-tailed Data
- Head class 개수는 많으나, tail class들의 샘플 수가 매우 부족한 경우
- 한계점 : 클래수가 불균형하면 모델이 head class에 편향, tail class에 대한 성능이 떨어짐
- Tail class 샘플 수 부족하면 tail class에 대한 분류 학습 힘들어짐



2. 세부 연구 내용

학습 포함 코드 목록 증가

- 메인 ipc 코드를 데이터 개수 50개 이상인 7개로 확대
- 4개만 포함한 모델에 비해 약 11% 정도의 정확도 감소를 보임
- 비록 정확도는 낮아졌으나, 더욱 실용적인 모델이라고 생각할 수 있음

```
epoch 10 batch id 1 loss 0.22292703390012146 train acc 0.96875
epoch 10 train acc 0.9791666666666666
100% 9/9 [00:04<00:00, 2.15it/s]
epoch 10 test acc 0.7384632616487455
```

테스트 데이터에 대해
74% 정도의 정확도를 보임

2. 세부 연구 내용

불균형 데이터 특성 개선

- 메인 ipc 코드별 데이터의 개수가 코드별로 상이하다는 점을 보완하고자 데이터를 무작위 복제
- 상위 4개 데이터 개수를 모두 200개로 설정
- 기존 85%에 비해 3% 증가된 정확도를 보였으나, 데이터를 복제했다는 점에서 오버피팅 문제 발생

```
100% ████████████████████████████████ 10/10 [00:06<00:00, 1.80it/s]
epoch 10 batch id 1 loss 0.06321728974580765 train acc 1.0
epoch 10 train acc 0.9796875
100% ████████████████████████████████ 4/4 [00:01<00:00, 3.74it/s]
epoch 10 test acc 0.87890625
```

테스트 데이터에 대해
88% 정도의 정확도를 보임

2. 세부 연구 내용

불균형 데이터 특성 개선

- 메인 ipc 코드를 최대한 사용하는 모델 고려 → 데이터 개수가 두자릿수인 모든 코드를 사용, 총 18개의 코드에 대한 학습 및 테스트 진행
- 데이터 불균형 해결을 위해 18개 코드 중 데이터 개수가 최소인 12개에 맞춰 총 234개의 데이터로 모델 구현
- 데이터의 개수가 매우 적기에 정확도가 낮아짐. 기존 상위 7개 포함 모델과 비교해 38% 감소 → 언더 샘플링 문제 발생

```
epoch 10 batch id 1 loss 2.3588058948516846 train acc 0.375
epoch 10 train acc 0.515625
100% 2/2 [00:01<00:00, 2.09it/s]
epoch 10 test acc 0.359375
```

테스트 데이터에 대해
36% 정도의 정확도를 보임

3. 연구 결과

연구 결과 분석

- 최종 채택 모델은 데이터 개수 50개 이상인 7개의 메인 ipc 코드 데이터를 활용한 모델 채택, 정확도 74%
- 4개 클래스 사용 모델은 활용 데이터가 758개 (전체 58.9%), 7개 클래스 모델은 955개(74.2%)로 감소한 정확도(약 11%)를 모델의 범용성으로 보완 가능하다고 판단

메인 ipc 코드	개수
G07F	262
G07D	196
G01G	181
G06Q	119
G05D	72
G16C	69
G05B	56
G07C	38
G08G	34
G09F	30

3. 연구 결과

의의

- 정확도를 높이기 위해서 클래스 개수를 낮추거나, 범용성을 높이기 위해 클래스 개수를 늘릴 수 있으나 두 가지 방법 모두 범용성의 감소와 정확도 감소라는 문제점 존재
- 따라서 적절한 클래수 개수를 활용해 범용성과 정확도를 모두 확보하는 것이 중요, 편향된 데이터 속에서 최대한 정확도를 확보하며 범용성을 높일 수 있는 모델 고안
- 특허 출원 빈도 기준 상위권 데이터에 대한 정확도 확보를 통해 특정 IPC/CPC 코드에 대한 식별 가능, 일부 분야에 대한 특허가 상대적으로 더 많이 출원된다는 실생활 상황 적용 가능

4. 한계점 및 제언

한계점

- 기존 Dataset 개수의 부족 : 총 개수가 1,286개로 이상적인 머신러닝 학습 Dataset 개수에 비해 많이 부족함
- Data 분포 특성의 문제 : 출원 빈도 상위 항목들에 대부분의 특허들이 몰려 있는 형태(long-tailed)로 학습 과정에서 상대적으로 출원빈도가 적은 특허들에 대한 학습이 부족, 정확도 하락과 연결

제언

- 더 큰 규모의 Dataset 확보 : 더욱 효율적인 머신러닝 학습을 위해 전체 데이터 개수가 많은 Dataset 확보 제안
- Data 분포 특성 해결 방안 : 주요 섹터의 특허 출원 빈도가 높은 것은 당연한 과정이므로 이를 해결할 수 있을만한 다른 전처리 과정이나 방법을 이용할 필요가 있음

인공지능과 자연어 처리를 통한 IPC 기술 분류

고려대학교 산업경영공학부 금강산, 김현진, 유선호

지식재산융합인재양성사업단 지식재산캠프

고려대학교 산업경영공학부