ISSN 2733-9696(온라인)
ISSN 2733-9572(인쇄본)

2021년 5호

GTC BRIEF

2021
OCTOBER
Vol.2 No.5

GTC BRIEF는 기후기술과 관련하여 시의성 있는 현안 및 동향정보를 알기 쉽게 정리한 자료임

GTC 녹색기술센터
GREEN TECHNOLOGY CENTER

ISSUE 02

2020년 국가연구개발과제의 기후기술 분류체계 기반 딥러닝 분류모델 적용 연구

주경원, 한수현 / 기술총괄부 | kwjoo@gtck.re.kr, sue@gtck.re.kr

하이라이트

- 딥러닝 기법을 활용한 AI 문서분류 알고리즘의 발전에 따라 국가 기후기술 연구개발과제의 기후기술 분류체계 기반 분류 알고리즘 개발 및 적용성 검토
- 딥러닝 모델은 텍스트 전처리, 워드 임베딩 벡터, 합성곱신경망(CNN), 장단기메모리(LSTM) 알고리즘을 활용하여 구축하였으며, 국가 연구개발과제의 과제명을 통해 기후기술 분류체계의 대분류, 중분류, 소분류를 예측
- 학습 및 테스트 데이터로 기존 기후기술 분류체계에 맞춰 분류된 4만여건의 국가 연구개발과제 정보를 활용하였으며, 대·중·소분류 별로 각각 약 90%, 78%, 70%의 분류 정확도를 나타냄
- 향후 데이터 증강 및 사전학습 활용도 확장을 통해 딥러닝 모델을 추가적으로 개선하고 국가 연구개발의 기후기술 분류검토 전문가 자문 시 보조자료로 활용

키워드

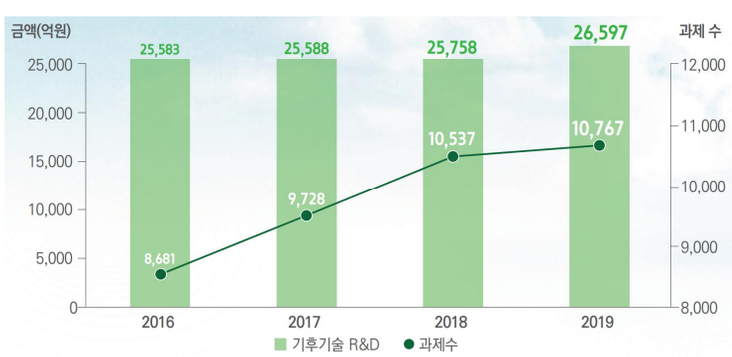
- 딥러닝, 다중분류, 합성곱신경망, 장단기메모리, 기후기술 분류체계

배경

기후기술 분류체계에 따른 국가 R&D 분류과정

- 기후기술 분류체계는 대분류-중분류-소분류 세 개의 계층적(hierarchical) 구조로 이루어져 있으며, 각각 3개, 14개, 45개의 기술분야로 정의됨
- 국가 연구개발과제 중 기후기술 R&D는 2016년도 8,681건에서 2019년 10,767건으로 매년 꾸준하게 성장하고 있으며, 투자액은 2019년 기준 약 2.7조원으로 국가 전체 R&D 중 12.9%를 차지[1)]

[그림 1] 국가 기후기술연구개발 투자금액 및 과제 수[1)]



※ 출처: 녹색기술센터 (2020), 2019 기후기술 국가연구개발사업 조사·분석 보고서

- 위와 같이, 기후기술 분류체계는 국가 연구개발의 기후기술 분야 별 연구개발 투자금액의 조사분석, 성과분석 등의 통계자료를 산출하는데 활용

- 녹색기술센터에서는 매년 한국과학기술기획평가원(KISTEP) 및 국가과학기술지식정보시스템(NTIS)을 통해 국가연구개발 정보를 이관받아 일차적으로 분류하고, 30~40명의 각 분야별 기술전문가의 자문을 통해 기술분류를 확정하고 있음

- 하지만 현 기술분류과정에서는 분야별 전문가의 경험적인 판단에 의존하고 있어 아래와 같은 문제점이 제기될 수 있음

  - '녹색기술센터 검토 – 전문가 자문 – 자문결과의 취합 및 통계량 검토'의 과정은 적지 않은 시간과 자문료가 소요되며, 보통 두 번의 라운드를 거치며 최초 연구개발정보 취득 후 분류결과를 확정하기까지 최소 3달 가량의 기간이 소요됨

  - 동일한 연구과제에 대해 각 전문가의 분류의견이 다를 경우 의사결정이 어려우며, 같은 전문가로부터도 동일한 연구과제에 대해 매년 다른 분류결과로 검토받는 경우가 다수 존재함

  - 자문을 수행하는 전문가가 자신의 전문분야에 투자되는 연구비가 과도하게 산출되어 공표될 것을 우려하며 의도적으로 기후기술에서 배제할 가능성

- 본 연구의 목적은 딥러닝 기반 AI 문서분류 알고리즘의 적용성을 검토하여 시의성 있는 국가 기후기술 연구개발의 통계자료 산출에 기여하고, 견고(robust)한 분류모델을 구현하기 위한 프로토타입을 개발하는 것임

AI 기반의 문서분류 연구동향

- 인터넷이 발달함에 따라 데이터의 양이 급격히 증가하고 있으며, 이에 따라 축적되는 문서들을 분석하기 위해 AI 기반의 다양한 텍스트 분석 방법들이 개발되고 있음. 본 연구에서는 문서 분류(classification)에 관한 연구동향을 분석

- 기계학습은 크게 지도학습(Supervised Learning), 비지도학습(Unsupervised Learning), 강화학습(Reinforcement Learning) 등으로 구분

  - 지도학습은 데이터별로 라벨*이 준비되어 있는 자료를 모형에 제공하여 훈련시키는 방법으로 크게 분류(classification) 및 회귀(regression)로 나뉘며, 비지도학습은 라벨링이 되어있지 않은 데이터를 통해 데이터가 가지고 있는 특성을 추출하거나 구조를 파악

    *각 데이터의 정답 (손글씨 이미지의 숫자, 사진 상 사물의 이름 및 위치, 뉴스의 카테고리 등)

  - 강화학습은 어떤 환경(Environment)이 주어졌을 경우 에이전트(Agent)가 현재의 상태(State)를 인식하고, 행동(Action)을 통해 보상(Reward)를 최대화하는 방향으로 학습하는 알고리즘으로 대표적인 예로 딥마인드(구글)의 알파고, 자동차의 자율주행 기술 등이 존재

- 감성분석(sentiment analysis)은 대표적인 문서분류 모델이며 상품이나 영화 등의 리뷰 문장과 그와 함께 매겨진 점수(별점)를 훈련자료로 사용
  - 고객 피드백, 콜센터 메시지 등과 같은 데이터를 분석하며, 외부적으로는 기업과 관련된 뉴스나 SNS 홍보물 등에 달린 댓글의 긍정, 부정을 판단하는데에 사용

- 스팸메일분류의 자동화는 가장 오랫동안 사용되어 온 AI 모델 중 하나이며, 과거에는 특정 키워드, 발신자 정보 등을 통해 필터링하였으나, 현재는 다양한 기계학습 및 방법론이 적용되어 높은 분류 성능을 나타냄

- 일반적으로 실생활에서 쓰이는 문서의 경우 사전학습(pre-trained)된 모델이 지속적으로 배포되어 왔으며, 최근에는 기술문헌과 같이 구조가 어렵고, 희소한 단어들이 많이 포함된 문장들에 대한 분석도 다양하게 이루어짐[2)3)]

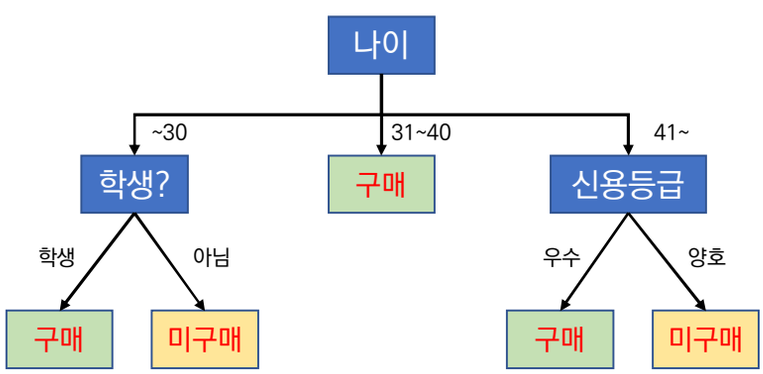
문서 분류를 위한 기계학습 방법론

의사결정트리 (Decision Tree)

- 의사결정트리를 이용한 기계학습 방법은 어떤 항목에 대한 관측값과 목표값을 연결시켜주는 예측 모델링 방법 중 하나이며, 분류와 회귀에서 모두 사용할 수 있기 때문에 CART(Classification And Regression Tree)라고도 함 (ex. 스무고개)

- 플로우 차트와 같이 분기점을 지날 때마다 특정 방향을 따라가는 구조를 가지고 있으며 각 잎(leaf) 노드는 클래스 레이블(결과)을 나타냄

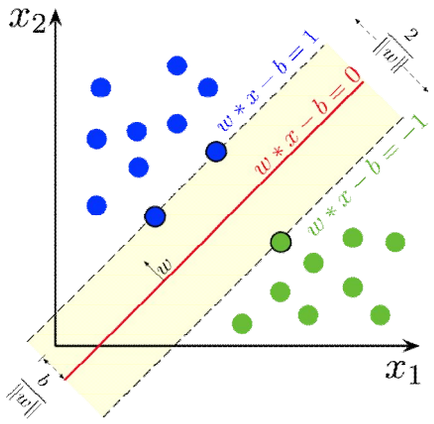
[그림 2] 고객의 정보를 바탕으로 구매여부를 예측하는 의사결정트리 예시



SVM (Support Vector Machine)

- SVM은 패턴 인식과 자료 분석을 위해 제안된 지도학습 모델이며, 두 개의 범주를 갖는 데이터의 집합에서 이진 선형 분류모델(경계)을 생성하여, 아래의 그림과 같이 가장 큰 폭을 갖는 경계선을 찾는 알고리즘임[4)]

[그림 3] SVM 경계 결정 알고리즘 모식도



※ 출처: https://en.wikipedia.org/wiki/Support-vector_machine (2021.09.06. 접근)

- 최근 신경망 기반의 딥러닝 방법이 개발되기 전에는 문서 분류작업에서 가장 우수한 성능을 나타내는 방법이었음[5)]

Naive Bayes

- Naive Bayes 자료들 사이의 독립(Independent)을 가정하는 베이즈 정리(Bayes' Theorem)를 적용한 확률이론에 기반한 분류 알고리즘으로 1950년대 이후 광범위하게 연구되어 적용되고 있으며 문서 분류에서는 스팸메일 필터링 알고리즘으로 주로 사용됨

- 예시로, 스팸메일(Spam)인지 정상 메일(Ham)인지 분류하는 Naive Bayes 추정식은 아래와 같으며 해당 계산결과의 부호에 따라 결과를 추정함

$$\log\left\{\frac{P(Ham)}{P(Spam)}\right\}+\sum_{i=1}^{m}\log\left\{\frac{P(w_i \mid Ham)}{P(w_i \mid Spam)}\right\}$$

- 여기서 $P(Ham)$, $P(Spam)$은 메일들이 정상 또는 스팸메일일 확률이며, m은 특정 메일 내 단어의 수, $P(w_i \mid Ham)$, $P(w_i \mid Spam)$은 단어 w_i가 정상 또는 스팸메일에서 등장하는 조건부확률

워드 임베딩 (Word Embedding)

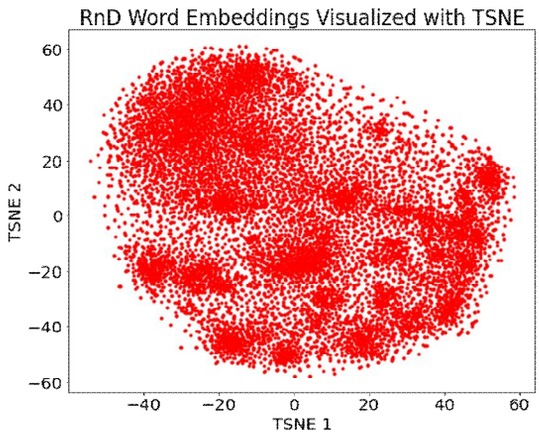
- 워드 임베딩은 컴퓨터가 자연어를 사람과 같이 이해할 수 있도록 단어를 벡터로 표현하는 방법으로 밀집 표현(dense representation)하는 과정을 통해 계산되며, 임베딩 벡터(embedding vector)라고도 부름

- 임베딩 벡터는 원-핫(one-hot) 인코딩 방법에 비해 낮은 메모리를 사용하며, 벡터의 값을 훈련 데이터로부터 학습하기 때문에 자연어 분석을 위한 딥러닝 모델링에서 첫 번째 레이어로 주로 사용

- 2016~2019년의 국가연구개발과제명의 워드 임베딩 벡터를 2차원으로 표현하면 아래의 그림과 같이 나타낼 수 있음

- 그림의 점은 국가 연구개발과제 단어사전의 각 단어들을 의미하며, 64차원으로 학습된 각 단어의 벡터공간을 t-SNE 방법을 통해 2차원으로 축소하여 X(TSNE1), Y(TSNE2)축으로 도시

[그림 4] 국가 연구개발과제 워드 임베딩 벡터

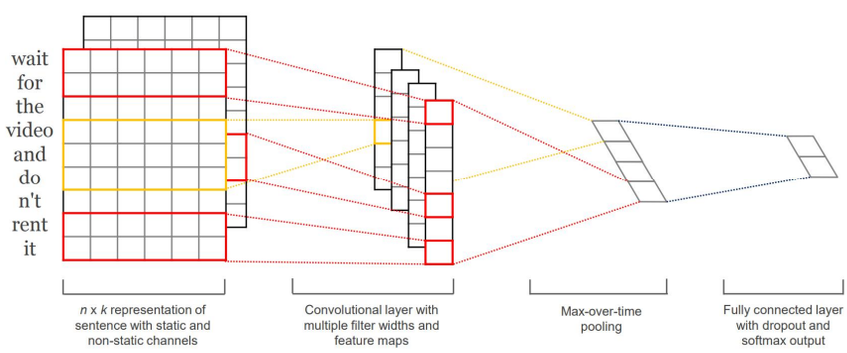


합성곱신경망 (CNN, Convolution Neural Network)

- CNN은 일반적으로 이미지 분류에 필수적인 딥러닝 방법으로 이미지의 영역 특성을 추출하는 데에 탁월하여 영상 및 동영상 인식, 추천 시스템, 영상 분류, 의료 영상 분석 등에 사용

- 텍스트를 워드 임베딩으로 표현할 시 이미지와 유사하게 한 축은 문장, 한 축은 임베딩 벡터로 표현할 수 있으므로 이미지 분류와 유사하게 적용할 수 있으며, 의미있는 특성들을 추출할 수 있음[5)]

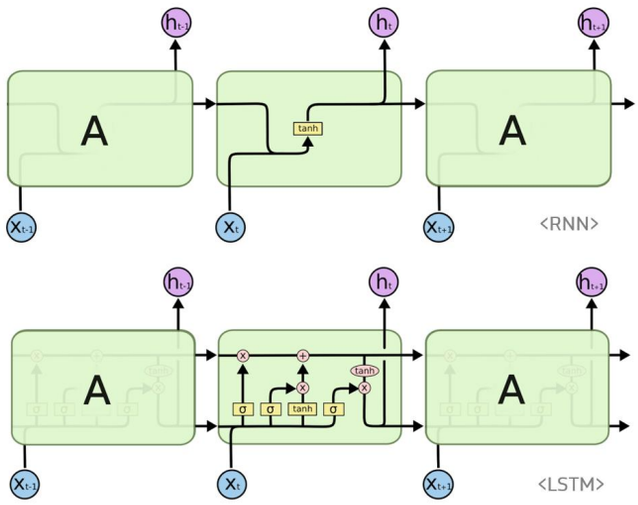
[그림 5] 문장 시퀀스에 대한 합성곱신경망 적용 예시[5)]



장단기메모리 (LSTM, Long Short-Term Memory)

- LSTM은 순환신경망(RNN)의 일종으로 입력 시퀀스가 길어질 경우 그래디언트 소실(vanishing gradient)를 해결하기 위해 고안된 방법으로 추가적인 은닉 층을 구성하여 의미있는 그래디언트가 유지될 수 있도록 개선된 방법임

[그림 6] 바닐라 RNN과 LSTM 구조 모식도



※ 출처: https://towardsdatascience.com/long-short-term-memory-networks-are-dying-whats-replacing-it -5ff3a99399fe (2021.09.06. 접근)

- 문서분류를 위한 순환신경망 구조는 여러개의 입력으로부터 하나의 출력을 가지는 Many-to-One 입출력 구조를 가지게 되며 마지막 토큰이 들어왔을 때 이를 완전연결층(fully-connected layer)로 넘겨 소프트맥스를 활용하여 어떤 범주에 속할 지 예측하는 순서로 계산

모델 적용 및 결과

모델 구조 및 데이터

- 국가연구개발의 기후기술 분류체계 기반 분류를 위한 기계학습 모형은 기존에 전문가 그룹을 통해 분류(라벨링)된 데이터를 활용하여 학습하는 모형으로 지도학습 방법에 속하며, 대·중·소분류에 따라 각각 3, 14, 45개의 분야를 예측하는 모델로 다중분류(Multi-class Classification)모형에 속함

- 2016~2019년 국가연구개발과제 중 기후기술로 분류된 39,713개의 데이터를 대상으로 적용

- 일반적으로 텍스트 전처리(preprocessing)시에는 한글 및 영문만을 대상으로 추출하지만, 연구개발과제 정보는 전문적인 기술문헌으로 단위(MW, Hz 등) 및 각 기술분야에서 통용되는 약어 등이 포함되어 있음

  - 따라서, 형태소 분석 후 제거하는 특수문자 등의 불용어(stopword)를 최소화하고, 출현 빈도가 낮은 한글, 영문의 고유명사도 워드 벡터에 포함함

- 훈련데이터와 테스트(검증)데이터는 8:2로 나누어 수행하였으며, 대·중·소분류 별로 데이터가 편중되지 않도록 계층적으로(stratified) 분할

  ※ 기계학습 모델 훈련 시 일반적으로 훈련/테스트 비율은 8:2로 분할하며 데이터의 수가 100,000개를 넘어가는 경우에는 훈련 데이터 비율을 상향함 (본 연구에서는 39,713개가 사용됨)

- 연구개발과제와 같은 기술문헌의 경우 고유한 단어가 많고 희소한 단어를 제외하기 어려워 입력계층의 차원이 큰 점을 고려하여 모델 구조는 임베딩 벡터-CNN(-LSTM)의 구조를 선택하였으며, 최적화 방법은 Adagrad의 문제점을 보완한 RMSProp을 사용함

- 각 모델별 하이퍼 파라미터를 설정하고 그에 따른 테스트 데이터를 기준으로 대분류, 중분류, 소분류별 정확도를 계산

분석결과

- 임베딩 벡터와 CNN을 활용한 모델 1~3번의 결과는 아래의 표와 같음
  - 1~3번 모델 별로 CNN필터의 수는 각각 16, 32, 64개를 적용하였으며, 3번 모델에는 2개의 레이어를 적용하고, 과적합을 방지하기 위해 50%의 dropout을 설정함

[표 1] 워드 임베딩 벡터와 CNN을 활용한 딥러닝 모델 구조 및 분류 결과 정확도

		Model		
		#1	#2	#3
Embedding Vector		64	64	64
CNN	Layers	1	1	2
	Filters	16	32	64
	Kernel Size	5	5	5
	Activation	relu	relu	relu
	MaxPooling	4	4	4
	Dropout	0.0	0.0	0.5
Batch		64	32	64
ReduceLROnPlateau		사용	사용	사용
Early stopping		5	5	5
정확도(%)	대분류	90.1	89.9	87.5
	중분류	77.4	77.6	78.1
	소분류	66.4	67.1	69.3

- 임베딩 벡터와 CNN을 활용한 모델 중 대분류 기준으로는 1번 모델(90.1%)이 가장 정확하며, 중분류와 소분류 기준으로는 3번 모델(78.1%, 69.3%)이 가장 높은 정확도를 나타냄

- 배치의 크기나 CNN의 레이어와 필터의 수는 정확도에 큰 영향을 미치지 않음

- 임베딩 벡터, CNN, LSTM을 활용한 모델 4~6번의 결과는 아래의 표와 같음
  - CNN 구조에서 4번 모델에는 3개의 레이어와 20%의 dropout을 적용하였으며, 5번 모델에는 40%의 dropout을 적용함
  - LSTM 구조에서 4번모델에는 16개의 레이어와 20%의 dropout을 적용하였으며 5, 6번 모델에는 32개의 레이어를 적용함

- 모델 컴파일 시 6번 모델에는 학습률을 조정하는 ReduceLROnPlatau을 적용하였으며 4, 5번 모델에는 학습률을 고정하여 학습함

[표 2] 워드 임베딩 벡터, CNN, LSTM을 활용한 딥러닝 모델 구조 및 분류 결과 정확도

		Model		
		#4	#5	#6
Embedding Vector		64	64	32
CNN	Layers	3	1	1
	Filter	32	32	32
	Kernel Size	5	5	5
	Activation	relu	relu	relu
	MaxPooling	4	4	4
	Dropout	0.2	0.4	0.0
LSTM	Layers	16	32	32
	Dropout	0.2	0.0	0.0
Batch		64	64	64
ReduceLROnPlateau		미사용	미사용	사용
Early stopping		5	3	5
정확도(%)	대분류	89.0	89.2	90.0
	중분류	77.9	77.3	78.3
	소분류	70.9	69.7	70.4

- 대분류, 중분류 정확도 기준으로는 모델 6번이 가장 우수(90.0%, 78.3%)하며, 소분류 기분으로는 모델 4번이 가장 우수(70.9%)하나 모델 간 큰 차이를 나타내지 않음

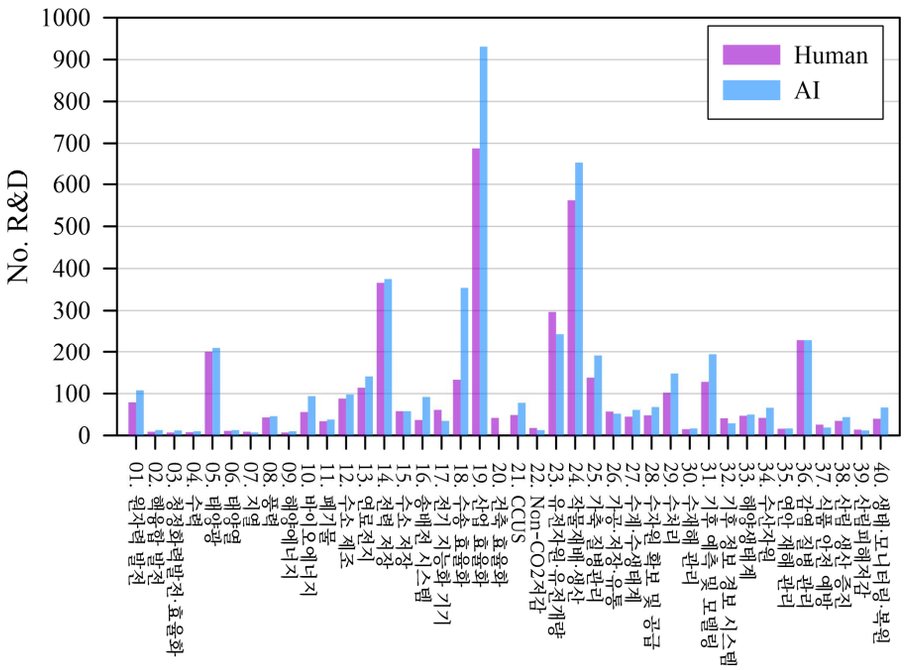
- 본 모델은 기존에 기후기술로 분류된 연구개발과제에 한해 적용한 결과이며, 전체 국가연구개발과제수인 연간 약 7만여개에 해당하는 데이터셋에 적용할 경우에는 소분류 기준 93.8%의 정확도를 나타냄

- 불균형(imbalance) 데이터의 성능 측정에 사용되는 f1-score(macro)의 경우 전 모델에서 0.70 내외의 수치를 나타내었으며, 2020년 신규 연구개발과제에 대한 AI 예측결과를 기술분류 전문가 자문 요청 시 보조자료로 제시함
  - 보조자료로 제공한 AI 모델의 예측결과는 기술분류 자문 시 긍정적인 피드백을 받았으며, 전문가가 분류한 약 70%의 연구개발과제는 전문가 검토결과와 동일하게 분류됨

2020년 신규과제 분석결과

- 본 모델의 적용성을 검토하기 위해 정확도 결과를 바탕으로 4번 모델과 6번 모델의 결과를 앙상블하여 훈련/테스트 과정에서 사용되지 않은 2020년 4,898개의 신규과제를 적용하여 소분류를 예측하였으며, 적응 및 감축 분야의 전문가 자문결과와 비교함
  ※ 그림 7의 AI 예측결과는 적용성 검토를 위해 개발된 모델의 예측결과로, 향후 사전학습 모델 활용과 앙상블 추론을 통한 최종 예측결과는 변경될 수 있음

[그림 7] 2020년 국가연구개발과제(신규)의 전문가그룹 및 딥러닝 모델 예측결과



- 전반적으로 전문가 분류결과와 유사한 예측결과를 나타내는 것을 확인하였으며, 유의하여 살펴봐야 할 점으로는 다음과 같은 특징이 있음
  - 전문가 그룹에 비해 AI는 더 많은 연구과제를 기후기술 분류체계 기반의 연구과제로 분류함
  - 19.산업효율화는 전문가 집단이 예측한 688개에 비해 높은 931개로 분류하였으며, 기존에 산업효율화로 분류된 연구개발과제의 단어 범위가 포괄적인 점이 영향을 미친 것으로 보임
  - 16.송배전시스템 또한 전문가 집단이 분류한 37개에 비해 높은 92개로 분류하였으며, 이는 송배전시스템 분야가 다루는 단어의 범위가 넓어 훈련과정에 영향을 미친 것으로 보임
  - 20.건축효율화의 경우 전문가 집단은 40개의 과제를 분류했지만 딥러닝 모델에서는 대부분 19.산업효율화로 분류하였는데, 산업효율화로 분류된 연구개발과제들이 건축분야의 단어 특성을 포함하고 있어 이와 같은 결과가 나타남

결론

- 본 연구에서는 AI, 딥러닝 분야에서 활발히 연구 및 활용되고 있는 문서분류 딥러닝 알고리즘을 활용하여 기후기술 국가연구개발과제의 한국어 말뭉치에 적용하였으며, 기후기술 분류체계 기반으로 정확도를 검토함
- 훈련 및 테스트 자료로 2016~2019년도의 국가 연구개발과제 중 기후기술로 분류된 39,713개의 연구개발과제를 활용하였으며, 이 중 80%를 훈련자료로 사용함
- 딥러닝 모델은 워드임베딩벡터, CNN, LSTM의 구조의 사용하였으며, 대분류, 중분류, 소분류 별 정확도는 각각 약 90%, 78%, 70%의 정확도를 나타냄

- 2020년 국가연구개발과제 중 신규과제에 대해 본 연구에서 학습시킨 모델 중 우수한 2개의 결과를 앙상블하여 예측한 결과와 전문가 집단의 분류결과와 비교한 결과 75.3%의 정확도를 나타내며, 아래와 같은 사항을 개선하여 추가적으로 모델의 성능을 제고할 필요가 있음
  - 소분류별 데이터 불균형을 고려하여 macro f1-score를 목적함수로 추가할 필요가 있으며, 각 소분류별 f1-score를 고려하여 여러 모델의 경험적 앙상블 추론 구현
  - 개선 방안으로 ①부족한 소분류에 대한 데이터 증강(augmentation), ②기술문헌을 해석하기 위한 사용자 사전 추가, ③텍스트 전처리 방법의 고도화, ④워드 임베딩 벡터의 사전학습 모델활용(koBert, koelectra 등)이 있으며, 향후 추가적인 연구를 통해 모델을 개선시킬 수 있을 것으로 기대함

- 본 모델을 통해 예측한 국가 연구개발과제의 기후기술 분야를 바탕으로 ①전문가 자문 과정에 소요되는 시간을 단축하고, ②예측 확률(softmax)이 높은 과제는 내부적으로 검토하여 전문가 자문에 소요되는 예산을 절감하고, ③전문가의 의견이 상충될 시 보조 지표로 사용할 수 있도록 활용 예정

참고문헌

1) 녹색기술센터 (2020), 2019 기후기술 국가연구개발사업 조사·분석 보고서

2) 황상흠, 김도현 (2020), 한국어 기술문서 분석을 위한 BERT 기반의 분류모델, 한국전자거래학회, Vol.25, No.1

3) KISTEP (2019), 기계학습 기반 바이오의료분야 과학기술정보데이터 분석활용 모형 고도화

4) Thorsten Joachims (1998), Text categorization with Support Vector Machines: Learning with many relevant features

5) Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., & Ngo, D. C. L. (2015), Text mining of news-headlines for FOREX market prediction: A Multi-layer Dimension Reduction Algorithm with semantics and sentiment. Expert Systems with Applications, 42(1), 306-324

6) Yoon Kim (2014), Convolutional Neural Networks for Sentence Classification (2014), arXiv:1408.5882 [cs.CL]

7) https://towardsdatascience.com/long-short-term-memory-networks-are-dying-whats-replacing-it-5ff3a99399fe (2021.09.06. 접근)

8) https://en.wikipedia.org/wiki/Support-vector_machine (2021.09.06. 접근)

본 내용은 녹색기술센터(GTC)의 주요사업(한수현, 안세진, 우아미, 주경원, 「기후기술 분류체계 기반 통계생산 및 국제확산」)으로 수행한 내용을 요약·정리한 것입니다.