컴퓨터학콜로키움 **(COSE-405) :: 12**주차 강재우 교수님

BioBERT : The first and the most widely used biomedical domain-specific language model

강의 : 2022-11-23 / 작성 날짜 : 2022-11-23

고려대학교 컴퓨터학과 2017320108

고재영

인류는 수천년을 거듭하여 매번 수많은 기술의 발전과 인식의 변화를 만들어 왔다. 인문학적인 분야, 의료생명공학적인 분야, 심지어는 예술적인 분야 등등 그 분야도 상당히 다양하지만, 그 중에서도 현재 가장 각광받고 있는 분야 중에 하나이며 타 분야와도 상당히 밀접한 연계를 이루는 부분을 말하자면 단연코 현대 IT의 꽃으로 볼 수 있는 인공지능 AI 및 기계학습 분야라고 해도 과언이 아니다. AI 기계학습의 강력한 점은 기술 그 자체 뿐만 아니라 건축/예술/생명공학 등 정말 다양한 타 분야에 접목시켜서 적용가능하다는 점일 것이다. 특히나 가장 화두가 되는 분야 단 세 가지를 꼽자면 건강의학, 경제, 그리고 생명 분야이다. 이번 시간에는 강재우 교수님께서 기계학습 및 데이터과학을 의료생명공학 쪽 분야와 결합하여 BioBERT에 대해 강연을 하셨다.

강재우 교수님께서 교수님의 연구실에서 해왔던 연구들을 소개해주시기에 앞서 먼저 필자와 같이 컴퓨터 학부의 공학도들은 다소 생소할 수 있는 의료 분야 개념들을 소개해주셨다. 항암 치료에 관한 약물 설계에 대한 내용으로 예시를 들어주셨다. 이 치료에 있어서 목표 인식의 단계를 4단계로 본다면 Target Identification, Lead Optimization, Pre-clinical, 그리고 Clinical Trials로 이루어져 있는데, 해당 과정에 있어서 기존의 의료진 및 연구진들에게 있어서 검색하는데에 큰 불편함이 있다는 문제점을 인식하셨다. 따라서 기존의 불편한 검색 서비스에 대해 보다 사용자 친화적인 방향으로 개선하기 위해 Biomedical Entity Search Tool로 불리는 BEST라는 검색 엔진을 연구하셨다. 검색 엔진 BEST에 대해 간단히 살펴보면, 기존에 존재하던 문헌에 대해서 index와 ranking에 대해 다루셨다. 기존 문헌에는 inverted indexing이 문서 단위로 존재하였는데 이 부분을 용이하게 바꾸고자 문서 내에 객체 단위로 indexing을

입힐 수 있게끔 논리적인 변환을 적용하였고, 이러한 변화를 적용하기 때문에 필연적으로 Ranking에 관해서도 조절을 적용했다고 하셨다. 그리고 BERT를 기반으로 사전 학습된 의료생명공학 분야의 텍스트 마이닝을 통한 언어 모델인 BioBERT를 개발하시며, 인공지능 질의응답 국제대회 BioASQ에서 우승을 차지하신 실적을 소개하셨다. 뿐만 아니라 코로나 시대에 굉장히 유명세를 탄 AstraZeneca 사와 협업을 통해 AstraZeneca-Sanger 드림 챌린지에서 약물과 표적단백질 상호작용에 대한 예측 및 다중표적 약물발굴에 대한 내용도 소개하셨다.

사실 강재우 교수님의 데이터 과학과 기계학습 학부 수업을 들었기 때문에 교수님의 의료생명공학 분야와 밀접한 연관성은 익히 알고 있었다. 하지만 교수님께서 해당 분야와 관련이 생기게 된 이야기를 하신 적이 있으셨는데, 우연한 기회로 당시 근무하시던 연구실 옆 연구실이 바이오 메디컬 분야였던 작은 계기로 시작되셨다는 부분이었다. 필자 또한 한 명의 공학도로서 단지 학부 때 배운 내용이 그 자체에 머무르지 않고 좀 더 실생활에서 다른 분야 혹은 미처 나와는 전혀 관련이 없을 것 같은 분야에 적용 가능할만한 우연한 관심을 가져야 할 것이다.