2023 강남대학교 데이터사이언스전공 졸업작품 발표회

트랜스포머 기반 영문 인공지능 PDF 논문 번역을 위한 기계번역 모델

함지율 이현도
강남대학교 데이터사이언스 전공
E-mail: hramsm@naver.com, hanse199911@naver.com

Introduction 서론

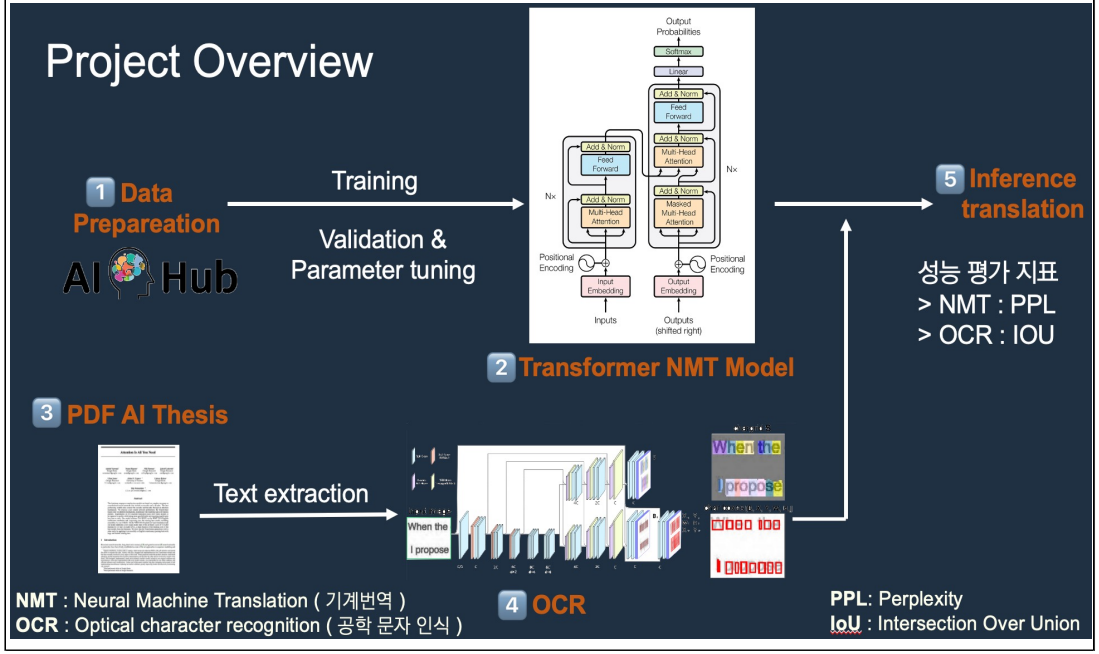
시중에 네이버 파파고, 구글 번역기 등 여러 번역기가 나와 있습니다. 하지만 특정 도메인 지식을 반영하여 번역하지 못한다는 단점을 파악하였습니다. 예를 들어, 이상 탐지 알고리즘 논문에 contamination을 **'오염'**이라고 번역합니다. 하지만 도메인 지식에 따르면 **'데이터의 불순도'**로 의역할 수 있습니다. 이러한 점을 반영하여 **(1) 영문 인공지능 논문을 한국어로 번역하여 학생들이 더욱 편하게 읽을 수 있도록 하고자** 주제를 선정하였습니다. 그리고 논문의 문장을 하나하나 스크롤 하여 번역기에 입력하는게 불편하였습니다. 저희는 PDF의 텍스트 또한 편리하게 **(2) PDF를 입력하면 이미지로 변환하고 변환된 이미지의 텍스트를 인식하여 한영 번역을 해주는 기계번역 모델(NMT)**를 구축하는 것을 목표로 프로젝트를 진행했습니다.

NMT : Neural Machine Translation (기계번역)

OCR : Optical Character Recognition (광학 문자 인식)

Method 방법론

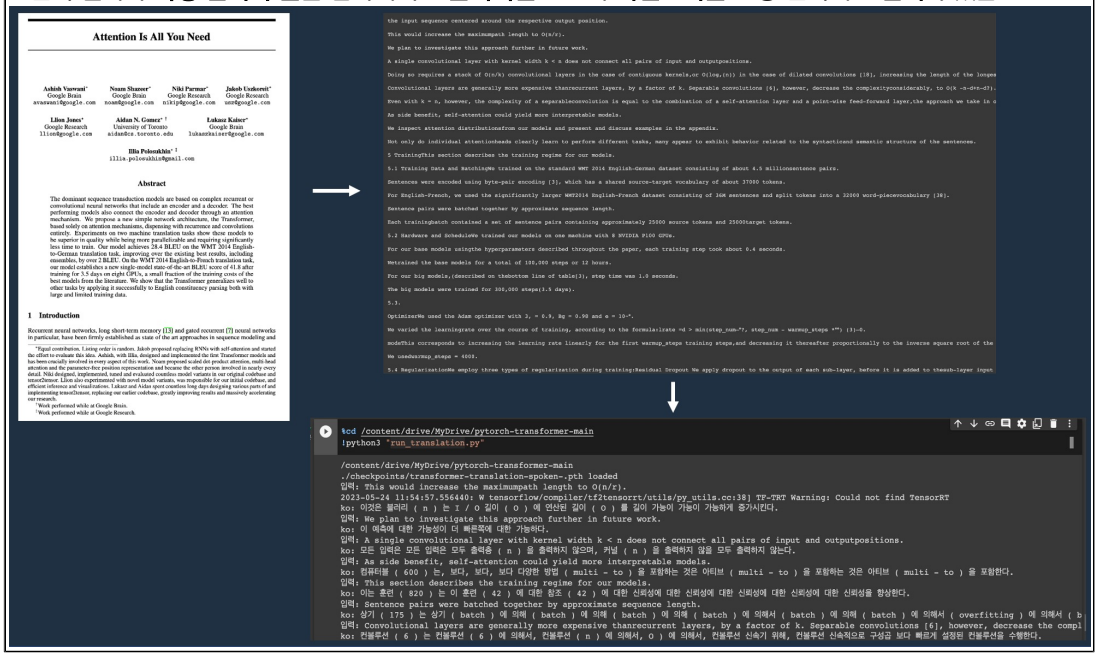
1) 수집한 데이터를 학습 데이터로 사용할 수 있도록 준비
2) Transformer를 통해 기술과학 논문 영한 번역 데이터와 일상 구어체 데이터를 통해 학습
3,4) OCR(Optical character recognition) 기술로 PDF의 텍스트를 추출
5) 추출한 텍스트를 transformer 모델을 통해 추론(번역)



Result 결과

1) Transformer를 소개한 'Attention Is All You Need' 논문 PDF에서 문서 추출
2) 문서 추출된 텍스트를 번역 모델에 입력값으로 넣어 번역

번역 결과가 **특정 단어 수준**을 잡아내지만 **문맥적인** 요소와 **자연스러운** 문장 번역에는 **한계**가 있음



Discussion 토의

- 성능 개선을 위한 추가적인 방안은 어떤 것이 있을까?
- 모델 학습시 하이퍼 파라미터를 어떻게 지정하는 것이 가장 효과적일까?
- 기계 번역에 State-of-the-art 성능을 보이는 다른 모델을 통해 학습하면 어떨까?

References 참고

1. Attention Is All You Need https://arxiv.org/abs/1706.03762
2. Transformer 코드 구현 깃허브 https://github.com/nawnoes/pytorch-transformer
3. OCR 기술 블로그 https://wandukong.tistory.com/9
4. Bert Tokenizer https://huggingface.co/bert-base-multilingual-cased