졸업작품 발표 요약서

팀 명	TP	담당 교수님	박동철교수님
팀 원	통계학과 1810704 김채현		
작 품 명	논문에 최적화된 번역 프로그램		
작품내용	최근 자연어 처리에 대한 관심이 커지면서, '번역' 분야는 빠르게 발전하고 있다. Attention Mechanism의 등장 이후, Transformer 모델을 시작으로 최근 BERT, GPT3 등의 언어 모델들은 모델 크기가 증가되어 상당히 정확한 결과를 보여준다. 하지만 Papago나 Google Translator와 같은 번역 사이트에서 논문 내용을 번역해보면, 컴퓨터과학 분야에서 통상적으로 영어 그 자체로 사용되는 단어들까지도 한글로 해석된 번역 결과를 보여준다. 가령 'epoch'와 같은 경우 컴퓨터과학 분야 논문에서는 epoch로 사용하여야 하지만, 번역 사이트를 통해 '시대'로 직역된다. 이에 논문 분야(인문학, 사회과학, 자연과학, 공학, 의약학, 농수해양학, 예술체육학)별로 최적화된 번역 프로그램을 구축하고자 한다. Naive Model은 Papago API를 이용하되, Wikipedia와 Oxford에서 Computer Science Jargon들을 크롤링해서 그 단어들에 한해 한국어로 직역하지 않고 영어 단어로 남겨둔다. Advanced Model은 Seq2Seq with Attention 모델을 사용한다. Korean Parallel Text Corpora (1,000 parallel sentences), Korean - English Parallel Corpus (700 training 700 test sentences), 그리고 ParaCrawl English - Korean (4,002,441 parallel sentences) Dataset을 사용하여 전처리한 뒤, Seq2Seq 모델에 Attention Mechanism을 사용한다. 이후 번역이 어색한 문장들에 대해 사용자들로부터 정확한 한국어-영어 번역 문장쌍을 입력받아 training dataset에 추가한다. 이 과정의 주 목표는 같은 단어이더라도 문맥상 한글이 더 자연스러운 경우에는 한글로, 영어가 더 자연스러운 경우에는 영어로 단어를 출력하도록 학습시키는 것이다.		
언어/환경	개발환경 : Windows 10. i7-5820K, RAM 16.0GB 개발언어 : Python 3.9.1 (필요 라이브러리: spaCy, NLTK) PyTorch Node.js 14.15.5 NPM 6.14.11 Html, Css, Javascript		
기타 건의안			