|  |  |
| --- | --- |
| 교육 제목 | 자연어 처리 개요 |
| 교육 일시 | 11월 25일(목) |
| 교육 장소 | 강의실 |
| **교육 내용** | |
|  | 자연어(Natural Language)란?   * 자연어 혹은 자연 언어는 사람들이 일상적으로 쓰는 언어를 인공적으로 만들어진 언어인 인공어와 구분하여 부르는 개념 * Natural Language Processing * 사람이 이해하는 자연어를 컴퓨터가 이해할 수 있는 값으로 바꾸는 과정(NLU) * 더 나아가 컴퓨터가 이해할 수 있는 값을 사람이 이해하도록 바꾸는 과정(NLG)   NLP는 무엇 때문에 어려울까?   * Ambiguty(모호성) : 언어는 마치 생명체와 같이 진화하며, 특히 효율성을 극대화 하는 방향으로 진화   따라서, 최대한 짧은 문장 내에 많은 정보를 담고자 한다.  정보량이 낮은 내용(context)은 생략  여기에서 모호함(ambiguity)이 발생  생략된 context를 인간은 효율적으로 채울 수 있지만, 기계는 이러한 task에 매우 취약하다.  한국어 NLP는 무엇이 더 어려울까?   * 교착어 라서   -> 접사 추가에 따른 의미가 파생 됨  -> 유연한 단어 순서 규칙  -> 모호한 띄어쓰기  -> 평서문과 의문문의 차이 부재  -> 주어 부재  -> 한자 기반의 언어  NLP Progect Workflow   * 문제정의 -> 데이터 수집 -> 데이터 전처리 및 분석 -> 알고리즘 적용(모델 설계) -> 평가 -> 배포   전처리 Workflow   * 데이터(corpus)수집 -> 정제 -> 레이블링 -> Tokenization -> Subword Segmentation -> Batchify * 정규식을 활용한 정제 * 정규식(Regular expression)을 활용하면 복잡한 규칙의 노이즈도 제거/치환 가능 * 코딩 없이 단순히 텍스트 에디터(Sublime Text, VSCode등)도 가능 * Interactive 노이즈 제거 과정   규칙에 의해 노이즈를 제거하기 때문에, 노이즈 전부를 제거하는 것은 어려움.  따라서, 반복적인 규칙 생성 및 적용 과정이 필요  끝이 없는 과정  노력과 품질사이의 trade-off, Sweet spot을 찾아야 함 |
|  |
| 전체 내용 |