|  |  |
| --- | --- |
| 교육 제목 | 자연어 처리 형태소 분류 |
| 교육 일시 | 11월 26일(금) |
| 교육 장소 | 강의실 |
| **교육 내용** | |
|  | * Labeling * Text Classification   Input : sentence  Output : class   * Token Classifiacation   Input : sentence  Output : tag for each token   * Sequence to sequence   Input : sentence  Output : sentence   * Tokenization * 한국어의 경우  1. 접사를 분리하여 희소성을 낮추고 2. 띄어쓰기를 통일하기 위해 tokenization을 수행  * 굉장히 많은 POS Tagger가 존재하는데  1. 전형적인 쉬운 문장(표준 문법을 따르며, 구조가 명확한문장)의 경우, 성능이 비슷함 2. 하지만 신조어나 고유명사를 처리하는 능력이 다름 3. 따라서, 주어진 문제에 맞는 정책을 가진 tagger를 선택하여 사용해야 함.  * Detokenization * 정제 :   Task와 언어 및 도메인에 따른 특성  풀고자 하는 문제의 특성에 따라 전처리 전략이 다름   * 끝이 없는 과정   노력과 품질 사이의 Trade-Off  Sweet spot을 찾아야함   * 분절   한국어의 경우 띄어쓰기 normalization을 위해 형태소 분석기 활용이 필요  Subword segmentation을 통해 좀 더 잘게 분절 할 수 있음   * 모두 비슷한 알고리즘을 사용하고 있으므로, 결국 데이터의 양과 품질이 좌우함   따라서 전처리 과정을 경시해서는 안됨 |
|  |
| 전체 내용 |