Labeling

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Preprocessing Workflow

데이터(코퍼스) 수집
- 구입, 외주
- 크롤링을 통한 수집

정제
- Task에 따른 노이즈 제거
- 인코딩 변환

레이블링(Optional)
- 문장마다 또는 단어마다 labeling 수행

Tokenization
- 형태소 분석기를 활용하여 분절 수행

Subword Segmentation (Optional)
- 단어보다 더 작은 의미 단위 추가 분절 수행

Batchify
- 사전 생성 및 word2index 맵핑 수행
- 효율화를 위한 전/후처리

Label

- Text Classification
  - INPUT: sentence
  - OUTPUT: class

- Token Classification
  - INPUT: sentence
  - OUTPUT: tag for each token → sequence

- Sequence-to-Sequence
  - INPUT: sentence
  - OUTPUT: sentence

Label Example

Sentence → Class

- TSV 형태의 하나의 파일
  - 각 row가 문장과 대응되는 레이블
  - 문장 column과 레이블 column 구성

Sentence → Sentence (Sequence)

- TSV 형태의 하나의 파일
  - 각 row가 대응되는 문장 쌍
  - 각 문장 별로 column을 구성

- 두 개 이상의 파일로 구성
  - 같은 순서의 row가 대응되는 문장 쌍
  - 한 문장당 여러 레이블이 존재 할 경우
    - e.g. 한국어 ←→ 영어 ←→ 중국어

Tip: 레이블링 직접 진행하기

- Human Labeling은 prototyping 시, 굉장히 강력한 도구 (두려워 하지 말자)

- 효율적인 레이블링 도구를 구성하자 (e.g. 엑셀)