Week1-4 과제

## 한국 스트리밍 서비스 (왓\*, 쿠\*플레이, 티\*)에서 시청자가 영화를 보고 남긴 리뷰를 긍정과 부정으로 나누어 볼 수 있는 대시보드를 만들려고 한다. **리뷰 긍부정 판별 모델**을 만들려고 할 때, NLP 리서처/엔지니어로서 어떤 의사 결정을 할 것인지 각 단계에 맞춰 작성해보자. (단, 수집된 리뷰 데이터의 개수가 1,000개 미만이라고 가정하자.)

대시 보드 예시.

| 긍정 | 부정 |
| --- | --- |
| ID:  REVIEW: | ID:  REVIEW: |
| ID: REVIEW: | ID: REVIEW: |

### 문제 정의

풀고자 하는 문제를 정의하세요. 또한 데이터 생성 시 고려해야할 사항이 있다면 무엇인지 설명하세요. (예, 만약 긍정 리뷰가 부정 리뷰보다 많은 경우 어떻게 해야 할까?, 길이가 정말 긴 리뷰는 어떻게 전처리 해야 할까?)

—- 시청자가 남긴 영화 리뷰를 분석하여 긍정과 부정으로 나누어 라벨링하는 모델

/ 긍정 리뷰가 부정 리뷰보다 많다면 타겟수의 불균형을 해결하기 위해 둘의 리뷰수를 최대한 1:1의 비율로 맞춰야 한다.

/ 길이가 정말 긴 리뷰가 있다면 불용어를 최대한 제거하고 그래도 너무 길고 의미 없는 내용이 많다면 아예 해당 리뷰를 제거하고 모델링을 한다.

### 오픈 데이터 셋 및 벤치 마크 조사

리뷰 긍부정 판별 모델에 사용할 수 있는 한국어 데이터 셋이 무엇이 있는지 찾아보고, 데이터 셋에 대한 설명과 링크를 정리하세요. 추가적으로 영어 데이터셋도 있다면 정리하세요.

—- 네이버 영화 리뷰

링크 : <https://github.com/e9t/nsmc/>

데이터는 id, document, label 로 이루어져 있고

id는 네이버에서 제공한 review id(숫자)

document는 사용자가 작성한 리뷰(한글)

label은 0은 부정, 1은 긍정으로 라벨링 되어 있다.

총200000개의 리뷰이고 train데이터는 150000개, test데이터는 50000개로 구성되어있다.

— 트위터 감성분석 데이터

링크 : <https://www.kaggle.com/jp797498e/twitter-entity-sentiment-analysis>

데이터는 Tweet ID, entity, sentiment, Tweet content로 되어있고

Tweet ID(숫자), entity(어떠한 주제)

sentiment는 긍정, 부정, 중립 3가지이다. (entity와 관련이 없으면 중립으로 간주)

Tweet content(영어)는 사용자의 트윗 내용이다.

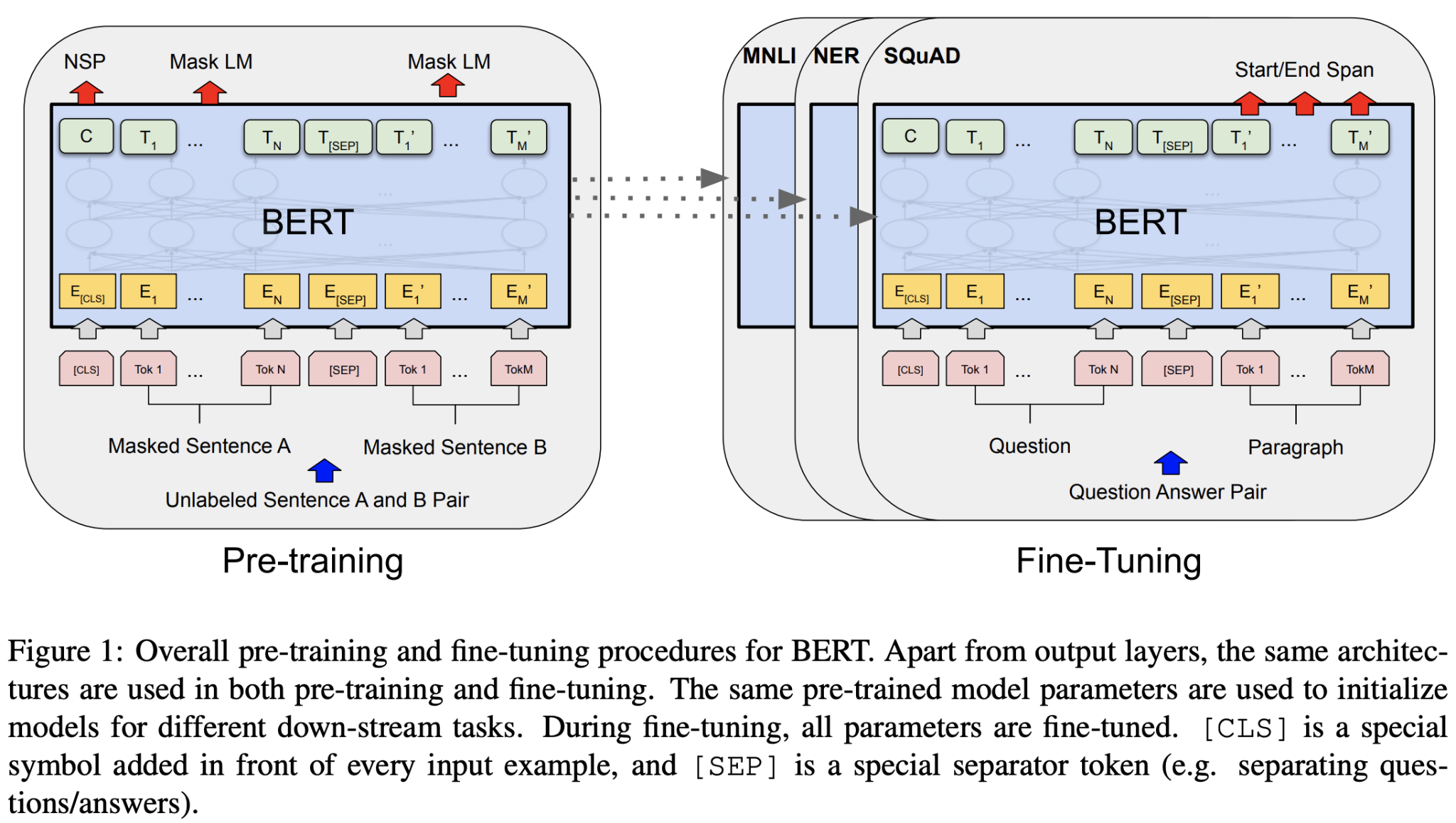
### 모델 조사

Paperswithcode(https://paperswithcode.com/)에서 리뷰 긍부정 판별 모델로 사용할 수 있는 SOTA 모델을 찾아보고 SOTA 모델의 구조에 대해 간략하게 설명하세요. (모델 논문을 자세히 읽지 않아도 괜찮습니다. 키워드 중심으로 설명해 주세요.)

—- SMART-RoBERTa Large

논문 : <https://paperswithcode.com/paper/smart-robust-and-efficient-fine-tuning-for>

RoBERTa(Robustly optimized BERT approach)



많은 양의 unlabeled corpus를 이용하여 언어 자체에 대해 배워나가는 pre-train단계와 특정 도메인의 task(down-stream task)를 학습하는 fine-tuning단계로 구성되며, 이 두 개의 단계를 차례로 거쳐 학습된다.

### 학습 방식

* 딥러닝 (Transfer Learning)  
  사전 학습된 모델을 활용하는 (transfer - learning)방식으로 학습하려고 합니다. 이 때 학습 과정을 간략하게 서술해주세요. (예. 데이터 전처리 → 사전 학습된 모델을 00에서 가져옴 → …)

데이터 전처리 → 전이학습된 모델을 github에서 가져옴(한글이면 https://github.com/SKTBrain/KoBERT) → 전이 학습된 모델을 나의 프로젝트에 맞게 재정의 → 쓰려는 데이터의 크기와 전이학습된 데이터와 쓰려는 데이터의 유사도에 따라 학습시킬지 안시킬지 결정 → 학습

* (Optional, 점수에 반영 X) 전통적인 방식  
  Transfer Learning 이전에 사용했던 방식 중 TF-IDF를 이용한 방법이 있습니다. TF-IDF를 이용한다고 했을 때, 학습 과정을 간략하게 서술해주세요.

—-

### 평가 방식

긍부정 예측 task에서 주로 사용하는 평가 지표를 최소 4개 조사하고 설명하세요.

—- 출처 : <https://www.koreascience.or.kr/article/JAKO201630762634226.pdf>

Precision : 긍정으로 예측한 경우 중 올바르게 긍정을 맞춘 비율

Recall : 실제 긍정인 것 중 올바르게 긍정을 맞춘 것의 비율

F1-score : precision과 recall의 조화평균

Accuracy : 옳게 맞춘 것을 전체 수로 나눈 값