1. 두 문장 관계 분류 task 소개

1.1 두 문장 관계 분류 task

> - 주어진 2개의 문장에 대해, 두 문장의 자연어 추론과 의미론적인 유사성을 측정하는 task

1.2 두 문장 관계 분류를 위한 데이터

> - **Natural Language Inference** (NLI)
>   - 언어모델이 자연어의 맥락을 이해할 수 있는지 검증하는 task
>   - 전체문장(Premise)과 가설문장(Hypothesis)을 Entailment(함의), Contradiction(모순), Neutral(중립)으로 분류
> - **Semantic text pair**
>   - 두 문장의 의미가 서로 같은 문장인지 검증하는 task

2. 두 문장 관계 분류 모델 학습

2.1 Information Retrieval Question and Answering (IRQA)

> - 기존 chatbot과의 차이점
>   - **Paraphrase Detection**
> - 1) 학습 데이터 구축
>   - 하나의 문장에 대한 similar_sents 리스트 생성
>   - similar_sents 리스트 안에 있는 문장은 유사한 문장
>   - 전체 sentence를 iteration 하면서 리스트 안에 없지만 유사한 문장 Top-N개의 리스트 생성 (non_similar_sents)
>   - non_similar_sents 리스트 안에 있는 문장은 유사하지 않은 문장
>   - **어려운 문제로 학습을 시키면 모델이 어려운 문제를 만났을 때 더 잘 맞힐 수 있다.**
> - 2) 유사도 비교 모델 학습
>   - 1번의 유사한 문장, 유사하지 않은 문장으로 모델 학습
> - 3) Chatbot
>   - 모든 sentence를 vector화 시킨다.
>   - Query vector와 모든 sentence와 유사도 비교
>   - Top-N개의 문장을 iteration 하면서 **2번에서 학습한 유사한 문장인지 판단**
>   - 유사한 문장일 경우 해당 Answer 출력 / 아닐 경우 다음 문장 확인

찐(?) 피어세션

> - (익효)
> - Data에 MASK를 씌우고 학습을 해볼 것
> - 종헌님 코드가 괜찮아 보인다.
>
> - (종헌)
>
> - Koelectra일 때 Unknown token을 모두 vocab에 추가할 때는 성능이 올라갔다.
> - XLMRobert에서는 좋지 않다...
>   - accuracy가 멈추는 구간이 많아진다.
>   - 학습에 방해가 된다고 생각해 쓰지 않는다.
> - 다양한 실험 끝에 tokenizer는 건드리지 않는걸로...
>
> - (재희)
>
> - entity 구분 token 추가
> - 2.4% 향상
> - ```
>   # 추천 방법
>   ENT1 </s> ENT2 </s> </s> <e1> ENT1 </e1> <e2> ENT2 </e2>
>   ```
> - (익효)
>
> - Embedding layer 실험해보지는 않았지만 피어세션 팀원들부터 나온 결과로는 오르지 않는다.
> - 아니였다...
> - 첫번째 문장 뒤에 "앞의 문장에서 a와 b는 어떤 관계야?"를 넣었을 때 성능이 좋지 않았다.
>
> - (종헌)
>
> - Entity를 앞에 붙여주는 것이 안 붙여주는 것보다 결과적으로 좋다.
>
> - (현규)
>
> - 이순신 [SEP] 조선 </s> </s> ...
> - 이순신 </s> 조선 </s> </s> ...
> - 위에서 아래로 바꾸었을 때 성능이 올라갈거라는 예상을 했지만 극락...
> - (익효) Entity 두 개를 연결해주는 영어 단어(SEPERATE)를 넣어주니 올라갔다.
> - (재희) 최대한 간결하게 하는 것이 좋을 거 같다.
> - _{이순신}은 <obj>조선</obj>중기 무신
>   이순신과 조선의 관계는 무엇인가?
> - 성능이 올랐다. (3%!!)
> - max_length를 150으로 늘렸을 때는 올리고 300은 떨어짐..
> - Question Type으로 진행할 예정
> - 번역을 사용했을 때 validation은 높았지면 결과로는 별로....
> - 학습이 고정되는 문제....가 자주 발생...
>   - (익효) learing rate를 계속 낮추면서 될때까지.. 또 안되면 SGD 무조건 될 때까지
>   - (종헌) 스케줄러 X 1e-5를 했을때 괜찮아짐
> - Overfitting 된다고 생각해 6 epoch
>   - (종헌) 5 epoch
>   - (재희) 5 epoch
> - 일반화 성능을 확인하고 그 다음에 ensemble하는 것이 좋다.
>   - 여러 실험을 해야 되기 때문에
> - (종헌)
>
> - 0.5 더해주는 것보다는 data 확률분포를 바탕으로 더해줘야 겠다.
>   - (익효) 0이 많이 틀리기 때문에 softmax 값을 취하고 10%를 더해주기 위해 0.5를 추가
>   - 0.7도 해보고 싶지만 제출 횟수가 너무 아깝다...
>   - 안올라갈수가 없다.
> - (현규)
>
> - Bucketting
> - get groups lenghts?
>   - 길이 비슷한 text끼리 group화를 시켜 padding을 최소화 시킴
>   - 처리해야 할 것들이 많다..
> - (재희)
>
> - Embedding layer 줄 수 있는 방법
>   - R-Bert
>   - 종헌님도 도전
> - (현규)
>
> - 데이터를 임의로 자르고 순서 바꾸기
>   - ex) 이순신은 조선 중기의 무신이다.
>   - ex) 조선 중기의 무신이다. 이순신은

새로 학습한 내용

시도한 것

> - wandb 적용
> - Entity 기준으로 양 옆 50만 보는 preprocessing 적용
>   - train 기준 200을 넘는 데이터가 거의 없다.(2-3개?)
> - Ensemble 적용(Hard voting)
>   - softvoting은 모델 저장이 안 돼 할 수가 없다....

TODO

> - Entity에 special token 추가해보기
> - 첫번째 문장 변경
>   - ex) ENT1, ENT2의 관계는?
> - 데이터를 임의로 자르고 순서 바꾸기
>   - ex) 이순신은 조선 중기의 무신이다.
>   - ex) 조선 중기의 무신이다. 이순신은