기술적인 도전

> - **점수 및 순위**
> - `accuracy` : 81.4%
> - `등수` : 5
>
> - **검증 전략**
>
>   - 9:1 비율로 train, validation set을 만들었습니다.
> - **사용 모델 아키텍처 및 하이퍼 파라미터**
>
> - **1) Main model** (78.9)
>
>   - `model` : xlm-roberta-large
>   - `preprocessing`
>   - entity 기준 양 옆 50 truncation
>   - ENT1 [SEP] ENT2 -> ENT1, ENT2 의 관계는?
>   - `loss` : cross-entropy
>   - `optimizer` : AdamW(weight_decay=1e-5)
>   - `sheduler` : lr_scheduler(warmup_steps=1000)
> - **2) Sub model 1** (78.7)
>
>   - `model` : xlm-roberta-large
>   - `preprocessing`
>   - entity 기준 양 옆 50 truncation
>   - sentence내의 ENT 강조 ex) ~ \<e1>ENT1\</e1> ~ \<e2>ENT2\</e2> ~
>   - `loss` : cross-entropy
>   - `optimizer` : AdamW(weight_decay=1e-5)
>   - `sheduler` : lr_scheduler(warmup_steps=1000)
> - **2) Sub model 2** (78.0)
>
>   - `model` : xlm-roberta-large
>   - `preprocessing`
>   - sentence max_length 100 -> 200
>   - ENT1 [SEP] ENT2 -> ENT1, ENT2 의 관계는?
>   - `loss` : cross-entropy
>   - `optimizer` : AdamW(weight_decay=1e-5)
>   - `sheduler` : lr_scheduler(warmup_steps=1000)
> - **2) Sub model 3** (77.4)
>
>   - `model` : xlm-roberta-large
>   - `preprocessing`
>   - entity 기준 양 옆 50 truncation
>   - `loss` : cross-entropy
>   - `optimizer` : AdamW(weight_decay=1e-5)
>   - `sheduler` : lr_scheduler(warmup_steps=1000)
> - 위 4개의 모델의 결과값 **hard voting** (3: 2: 2: 2)
>
> - **기타 시도**
>
>   - sentence에 랜덤하게 mask 생성
>   - ex) 이순신은 조선 중기의 무신이다.
>   - ex) 이순신은 \<mask> 중기의 무신이다.
>   - Hugging face의 여러 모델 사용
>   - KoElectra
>   - Bert
> - `IDEA`
>
>   - **1) sentence max_length 100 -> 200**
>   - sentence 길이에 대한 분포가 나와 있는 토론글이 있었는데 99%에 해당하는 값이 190이라고 나와 있어 200으로 바꿨습니다.
>   - 100으로 했을 경우에 sentence내에서 entity가 잘리는 경우가 있어 늘려야겠다는 생각을 했습니다.
>   - 결과적으로는 성능이 향상되었습니다.
>   - **2) entity 기준 양 옆 50 truncation**
>   - 학습 중 길이가 긴 data를 살펴보았는데 굳이 모든 단어를 봐야 하는지 의문이 들었습니다.
>   - 그래서 entity 기준으로 양옆으로 50 정도만 보아도 저는 관계를 유추할 수 있다는 생각이 들어 적용을 해보았습니다.
>   - 결과적으로 가장 성능이 급격하게 올라갔습니다.
>   - **3) ENT1 [SEP] ENT2 -> ENT1, ENT2 의 관계는?**
>   - 앞의 문장을 사람이 이해할 수 있는 문장으로 바꾸면 모델도 잘 학습하지 않을까라는 의문이 들었습니다.
>   - 관계 추출 문제이므로 위와 같은 문장으로 바꿔 적용했습니다.
>   - 기존의 [SEP] 보다는 소량으로 올라갔습니다.
>   - **4) sentence내의 ENT 강조**
>   - sentence내 entity의 단어가 여러 번 등장하는 문장들이 있어 모델이 봐야 하는 entity를 알려주면 어떨까 하는 생각이 들었습니다.
>   - 다음과 같이 적용을 했습니다. ex) ~ \<e1>ENT1\</e1> ~ \<e2>ENT2\</e2> ~
>   - Special token에 추가했을 때는 성능이 떨어졌습니다.
>   - Special token에 추가하지 않고 학습을 돌릴 때는 전반적으로 올라갔습니다.

교훈, 아쉬운 점, 계획

> - 어려운 문제를 해결할 때 사람들과의 **커뮤니케이션**이 정말 중요하다는 것을 알게 되었습니다. 사람마다 생각하는 방향이 다르니 여러 아이디어를 얻을 수 있었습니다. 사실 STAGE 3, 4에 같이 할 팀원들과 더 많은 소통을 하였고 여러 아이디어를 공유해주어 저에게 많은 자극이 되었습니다. 좋은 아이디어들이 많았지만 모두 제 것으로 소화하지 못해 아쉬웠습니다.
>   - 인상깊었던 아이디어는 다음과 같습니다.
>   - 주어진 dataset을 ner dataset으로 만드는 방법
>   - dataset에 랜덤하게 MASK 씌운 후 학습
>   - R-BERT
>   - Model 뒷 단에 LSTM 추가해 학습
> - P stage 1에서는 스스로 근거가 없어도 생각나는 아이디어가 있으면 바로바로 적용해 실패했던 경험이 있었습니다. 이번 stage에서는 **적용해보기 전에 의문을 가지게 되었고** 충분히 적용한다면 **좋은 아이디어라고 생각이 될 때 코딩**을 했습니다. 그로 인해 P stage 1 보다는 효과가 있었던 여러 아이디어를 적용할 수 있었습니다.
> - `TEAM-IKYO`에서 좋은 코드들을 공유해주었지만 제 것으로 소화하지 못한다면 쓰지 않았습니다. 아직 많이 부족함을 다시 깨닫게 되었습니다. 좋은 팀원들이 있어 참 감사하게 생각합니다.
> - 주어진 baseline을 새로 저만의 baseline을 만들었지만 하루 만에 실수로 삭제해.. 포기하고 다시 만들지 않아 조금 아쉬운 마음이 있습니다. 백업도 철저하게 해야겠다는 생각이 들었습니다.
> - Stage 1에서 계획했던 **EDA**를 다양한 관점으로 적용해 보아 여러 좋은 아이디어를 낼 수 있었습니다. 또한 **Wandb**를 통해 학습했던 log에 대해 관찰할 수 있었습니다. argparse는 적용해 보지 못해 다음 stage에는 적용할 예정입니다.