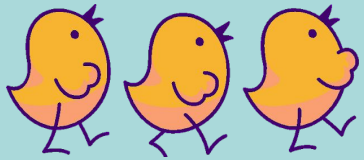


LangCon 2021

KLUE:

Korean Language Understanding Evaluation

Upstage 문지형

KLUE at KLUE Paper Day

Language Conference

KLUE Paper Day

KLUE _ Korean Language Understanding Evaluation

2021.06.01

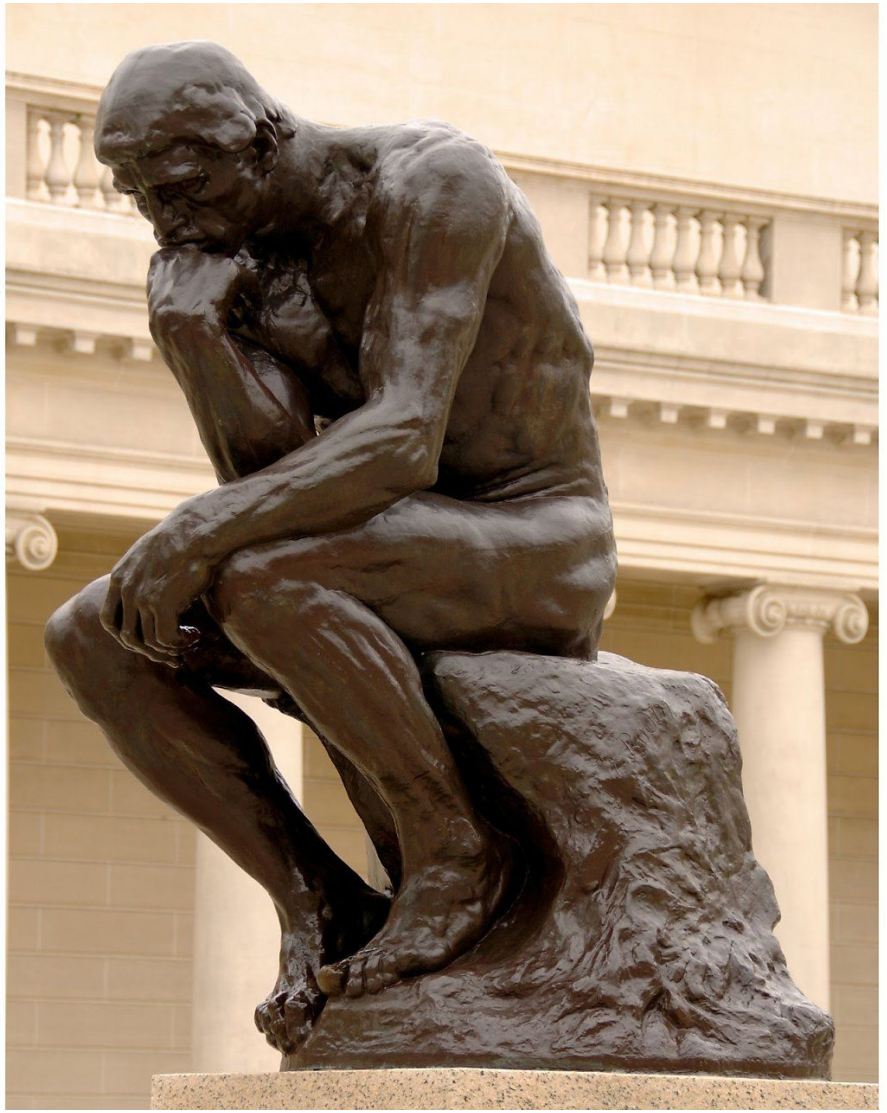
- Introduction
- Source Corpora
- Tasks
- Language Models
- Ethical Considerations
- Discussion
- Leaderboard

https://www.youtube.com/watch?v=0ySceOe7Rws

이미 많은 내용이 (심지어 public 하게) 공유되었다

Language Conference

페이퍼 소개를 또 하자니 중복 컨텐츠 발표란 내 사전에 없는데...



Paper Day에서처럼 페이퍼 소개를 한다고 하더라도 주어진 발표 시간은 30분 밖에 없어서 촉박할 것 같다...

처음 공유하는 내용이면서도 LangCon 취지에 맞고 유익한 주제는 없을까...

출처: https://post.naver.com/viewer/postView.nhn?volumeNo=13840762&memberNo=12200947

1. KLUE Overview
2. KLUE 구축 과정
    1. 태스크 선정 배경
    2. 널리 쓰이는 데이터를 위해
3. Baseline 코드 소개
4. KLUE Leaderboard
5. KLUE 이후의 데이터(에 대한 개인적인 바람)

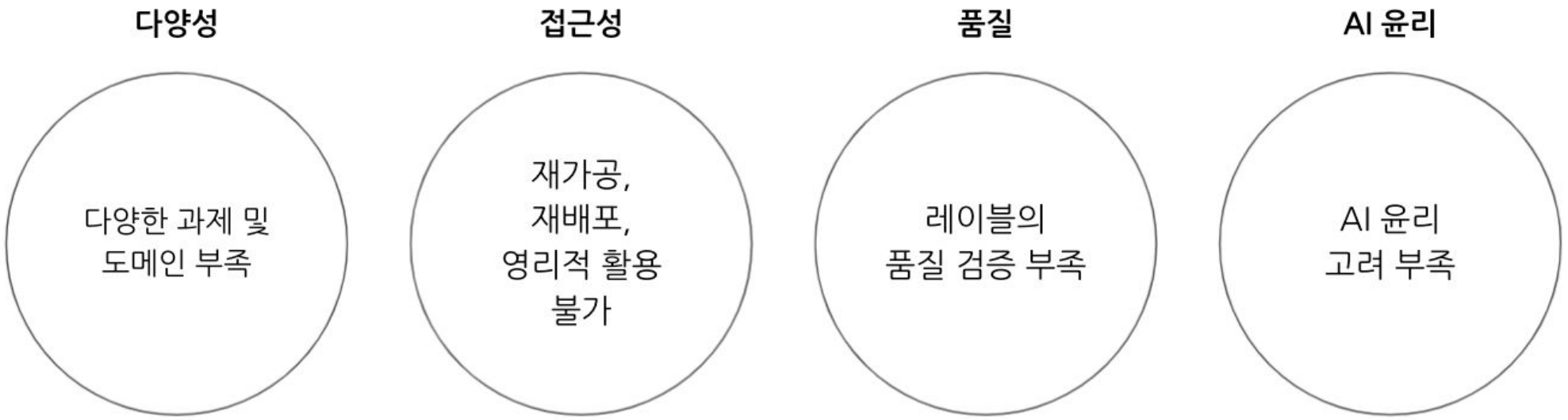
KLUE Overview

LangCon 2021

KLUE 제작 목적

Language Conference

- **자연어 이해 (NLU) 모델의 평가**
  - AI 가 사람의 언어를 얼마나 잘 이해하는지 평가하려면 다양한 과제에서 평가되어야 함
- **자연어 이해 (NLU) 모델 발전의 견인**
  - 공정한 자연어 이해 (생성) 모델 비교의 토대
- **기존 한국어 데이터셋의 한계 극복**



KLUE Task

Language Conference

Name	Type	Format	Eval. Metric	# Classes	|Train|	|Dev|	|Test|	Source	Style
YNAT	문장 주제 분류	Single Sentence Classification	Macro F1	7	45k	9k	9k	뉴스 (헤드라인)	문어체
KLUE-STS	문장 유사도	Sentence Pair Regression /Classification	Pearson's r F1	[0,5] 2	11k	0.5k	1k	뉴스, 리뷰, 스마트홈 쿼리	문어체, 구어체
KLUE-NLI	자연어 추론	Sentence Pair Classification	Accuracy	3	25k	3k	3k	뉴스, 리뷰, 위키피디아	문어체, 구어체
KLUE-NER	개체명 인식	Sequence Tagging	Entity-level Macro F1 Character-level Macro F1	6 12	21k	5k	5k	뉴스, 리뷰	문어체, 구어체
KLUE-RE	관계 추출	Single Sentence Classification	Micro F1 AUPRC	29 30	32k	8k	8k	뉴스, 위키피디아	문어체
KLUE-DP	의존 구문 분석	Sequence Tagging	UAC LAC	단어 수 38	10k	2k	2.5k	뉴스, 리뷰	문어체, 구어체
KLUE-MRC	기계 독해	Span Prediction	EM ROUGE-W	2	12k	8k	9k	뉴스, 위키피디아	문어체
WoS	대화 상태 추적	Slot-Value Prediction	JGA Slot Micro F1	(45)	8k	1k	1k	목적형 대화	구어체

KLUE 구축 과정

어떤 태스크를 선정해야 할까?

Language Conference

- **Ranking Criteria**
  - NLU 필수 태스크인가?
  - 다른 Benchmark 와의 차별성이 있는가?
  - 한국어만의 특성이 고려되어야 하는 태스크인가?
  - 회사에서 관심있을 법한 태스크인가?
  - 연구자에게 관심있을 법한 태스크인가?
  - 문제 정의가 견고한가?
  - 저작권 이슈가 없는가?
  - 구축 난이도는 적당한가? (제한된 시간 내에 프로젝트에 참여한 인력으로 만들 수 있는가?)

어떤 태스크를 선정해야 할까?

Language Conference

	Task	Task format	Ranking criteria									
			NLU 필수 task	다른 benchmark 와의 차별성	한국어 특성 고려	industrial needs	academic needs	문제 정의의 solidity	License free	구축 난이도	총점	투표인원	평점
			3	2	1	2	2	0.5	0.5	1			
	예시	예시									16	4	4.00
Sem	Dialogue State Tracking (Task Oriented Dialog)	지형, 지윤									45.5	8	5.69
Sem	Paraphrase	sentence 1 | sentence 2 | class									31	7	4.43
Sem	Hatespeech Detection	single sentence | hate label (hate, offensive, none)									9.5	4	2.38
Sem	Speech Act Classification	sentence | speech act class									10	4	2.50
Sem	ASR error을 포함한 text의 SLU	sentence | class									24.5	7	3.50
Sem	Open Domain Dialogue Understanding										21	5	4.20
Sem	STS	sentence 1 | sentence 2 | score (0-5)									14.5	5	2.90
Sem	NLI	sentence 1 | sentence 2 | class (contradicton, entailement, neutral)									22.5	7	3.21
Sem	Event Detection	sentence | event class									8	3	2.67
Sem	높임말 classification	sentence | class									11	7	1.57
Sem	사투리 classification	sentence | class									13	7	1.86
Sem	Relation Extraction	subject | object | relation | NE tagged single sentence									27	5	5.40
Sem	Intent classification	sentence | class									32	7	4.57
Sem	(Aspect) Sentiment Classification										32	7	4.57
Sem	Sarcasm Detection	sentence | binary class									19	6	3.17
Sem	News classification	sentence | class (news cateogry)									1	2	0.50
Sem	Temporal QA										8	2	4.00
Sem	non-answerable question classification	passage | question | answer (yes/no/I don't know)									14	4	3.50
Sem	**Machine Reading Comprehension**	지형,									34.5	6	5.75
Sem	Mathematical Reasoning	passage | question | answer									22	6	3.67
Sem	Common Sense QA	sentence 1 | relation (cause, result) | multiple choice									27	6	4.50
Sem	extractive summarization										12.5	4	3.13
Sem	Slot filling										11	4	2.75
Sy, Sem	Word sense disambiguation										12	5	2.40
Sy	형태분석(POS tagging)										20.5	5	4.10
Sy	구문 분석(Dependency Parsing) + 형태분석 분석 결과 포함	Conll 형태 분석 결과와 구문 분석 결과 범기(sejong tags), 아래 예시는 영문 U 1 Mary _ _ _ _ 2 nsubj 2:nsubj 2 won _ _ _ _ 0 root 0:root _ 3 silver _ _ _ _ 2 obj 2:obj _ 4 and _ _ _ _ 5 cc E5.1:cc _ 5 Sue _ _ _ _ 2 conj E5.1:nsubj 5.1 _ _ _ _ _ 2 conj 2:conj _ 6 bronze _ _ _ _ 5 orphan E5.1:dob									33	8	4.13
Sy	어휘의미 분석										8	4	2.00
Sy, Sem	상호참조 해결(Coreference resolution)										28	7	4.00
Sy, Sem	개체명 분석(NER, Named Entity Recognition)	지형, 지윤									33	8	4.13
Sy, Sem	의미역 분석(Semantic Role Labeling)										2.5	2	1.25
Sy, Sem	언어 수용성(Linguistic Acceptability)										10.5	4	2.63
Sy, Sem	무형 대용어 복원 (Anaphora resolution)										12	5	2.40

선정되지 못했던 태스크

- Hate-speech Detection
- Sarcasm Detection
- SLU (Spoken Language Understanding)
- 높임말/사투리 Classification
- Extractive Summarization
- Temporal QA
- Mathematical Reasoning
- Common Sense QA
- Coreference Resolution
- 무형대용어 복원

...

선정된 8개 Task를 어떻게 잘 만들 수 있을까?

Language Conference

- 사용가능한 Source Corpora 탐색 및 수집

담당자	크롤링 완료	크롤링 우선순위	코퍼스
		1	연합뉴스 (뉴스제목)
	(진행중)	1	위키트리 (본문)
	O	1	wikipedia
	O	1	정책브리핑
	O	2	네이버 영화 리뷰
	O	2	airbnb
	O	3	위키뉴스
		Generated	ParaKQC
		Generated	ToD
	O	2	리브레위키
		2	위키문헌
	O	2	위키책
		2	판결문
	변환	2	공유마당 시
		3	뉴스페퍼민트 (뉴스기사)
			annotate / create
12/12(토)까지 위 코퍼스 크롤링. 화이팅!!!			태깅플랫폼
		3	국립국어원 표준국어대사전
		3	우리말샘
		3	국회 회의록
	변환	999	카이스트 북 코퍼스
		999	UD corpus
		999	제타위키
		999	위키배움터
		3	네이버 연예 뉴스 리뷰
		3	네이버 쇼핑 리뷰
		999	서울 열린데이터광장
		999	공유만료저작물
		999	공공누리
		999	K-QuAD
		999	Korean Contemporary Corpus of Written Sentences
		999	코로나 시국의 일기 데이터
		999	책/소설 등
	변환	999	korean_parallel_data

선정된 8개 Task를 어떻게 잘 만들 수 있을까?

Language Conference

- **선정된 태스크에 분배**
  - 문어체와 구어체가 섞여있으면 좋은 태스크인가?
    - DP, NER, STS, NLI, RE, SSC
  - 구어체로 구축이 가능한 태스크인가?
    - DP, NER, STS, NLI

코퍼스	DP	NER	RE	SSC	STS	NLI	ToD	QA
연합뉴스 (뉴스제목)				o				
위키트리 (본문)	o	o	o			o		
wikipedia			o			o		o
정책브리핑			o		o	o		
네이버 영화 리뷰		o				o		
airbnb	o				o	o		
위키뉴스						o		
ParaKQC					o			
ToD							o	
리브레위키								
위키문헌								
위키책								
판결문								
공유마당 시								
뉴스페퍼민트 (뉴스기사)								

선정된 8개 Task를 어떻게 잘 만들 수 있을까?

Language Conference

- **선정된 태스크에 분배**
  - 문어/구어 Corpora 중에 해당 태스크에 적합한 코퍼스는 무엇인가?
    - 네이버 영화 리뷰: 영화 도메인에 대한 텍스트이기 때문에 영화 감독, 영화명, 배우명 등 공인에 해당하는 인물명과 개체명이 많은 편 → NER
    - 위키트리: CC-BY-SA 로 사용가능한 뉴스 코퍼스이지만 뉴스 컨텐츠 퀄리티 이슈가 존재 (click bait 등). 단 문장 단위로 분리하고 random sampling한 경우 문장 자체에 대한 이슈는 적음 → DP, NER, RE, NLI
    - 위키피디아: MRC를 제작하기 위해서는 문단 단위의 텍스트가 꼭 필요 (차별성을 위해 추후 뉴스 코퍼스 추가)

코퍼스	DP	NER	RE	SSC	STS	NLI	ToD	QA
연합뉴스 (뉴스제목)				o				
위키트리 (본문)	o	o	o			o		
wikipedia			o			o		o
정책브리핑			o		o	o		
네이버 영화 리뷰		o				o		
airbnb	o				o	o		
위키뉴스						o		
ParaKQC					o			
ToD							o	
리브레위키								
위키문헌								
위키책								
판결문								
공유마당 시								
뉴스페퍼민트 (뉴스기사)								

선정된 8개 Task를 어떻게 잘 만들 수 있을까?

Language Conference

- **선정된 태스크에 분배**
  - 수집한 코퍼스의 수량
    - 위키트리: 양이 많았기 때문에 많은 태스크에서 사용가능
    - 위키피디아/정책브리핑: 퀄리티 대비 양이 적어서 꼭 필요한 태스크에 우선 분배
  - General Domain 에 적합한가?
    - 판결문 X

코퍼스	DP	NER	RE	SSC	STS	NLI	ToD	QA
연합뉴스 (뉴스제목)				o				
위키트리 (본문)	o	o	o			o		
wikipedia			o			o		o
정책브리핑			o		o	o		
네이버 영화 리뷰		o				o		
airbnb	o				o	o		
위키뉴스						o		
ParaKQC					o			
ToD							o	
리브레위키								
위키문헌								
위키책								
판결문								
공유마당 시								
뉴스페퍼민트 (뉴스기사)								

태스크에 적합한 데이터 수량은 어느 정도인가?

Language Conference

다다익선

태스크에 적합한 데이터 수량은 어느 정도인가?

~~다다익선~~

충분한 시간과 자금과 사람이 있다면...

태스크에 적합한 데이터 수량은 어느 정도인가?

Language Conference

- **각 태스크에서 가장 작은 단위의 데이터를 구축할 때 드는 금액은 어느 정도인가?**
  - (POS) + DP: ~~겁나~~ 비쌈
- **데이터 구축을 위해 사용가능한 금액은 얼마인가?**

- **각 태스크에서 가장 작은 단위의 데이터를 구축할 때 드는 시간은 어느 정도인가?**
  - MRC: 요구 조건이 빡세서 은근히 오래 걸림. 지문 독해도 필요
- **데이터 구축에 쓸 수 있는 시간은 어느 정도인가?**

- **모델이 태스크를 풀기 위해 필요한 최소 수량은 어느 정도인가?**
  - 모델 성능이 안 나오는 이유가 데이터 부족이 아니길

각 태스크에서 추구해야하는 구축 방향은 무엇인가?

Language Conference

- **YNAT**
  - 본문을 보고 판단되는 predefined 뉴스 카테고리의 매칭 오류를 최소화
- **STS**
  - STS-b 의 구축방식을 최대한 Follow
  - 두 문장이 자연스러운 문장일 때 의미적으로 유사한 문장쌍과 그렇지 않은 문장쌍 pair를 자동으로 매칭시켜서 구축 시간을 단축하고 퀄리티를 향상
- **NLI**
  - 알려진 annotation artifact를 최소화하여 구축하는 것
- **NER/DP**
  - 첫 구어 데이터셋 (Source Corpora 선정에서 만족)
- **RE**
  - 충분한 수량의 데이터셋
    - Distant Supervision 및 NER 모델 사용

각 태스크에서 추구해야하는 구축 방향은 무엇인가?

Language Conference

- MRC
  - 이미 공개된 한국어 MRC 데이터셋에서 다루지 않았던 질문 유형 추가
  - 알려진 annotation artifact를 최소화하여 구축하는 것
    - 첫 문장에 정답이 있는 경우가 많음
    - multi-hop question 일지라도 한 문장만 보고도 맞출 수 있는 질문인 경우가 많음
- WoS
  - WoZ Setting에서 알려진 구축 방식의 한계를 극복한 방식으로 구축하는 것
    - Self-Dialog: 비용 감축
    - Dropdown menu: 에러 발생 최소화

어노테이션 퀄리티 향상

Language Conference

- 가이드라인의 모호함을 최대한 제거
  - 여러 번의 iteration
- 꼼꼼한 검수
- 제작자의 의도를 명료하게 전달
- 전문 지식이 필요한 경우 annotator 훈련을 충분히 시도

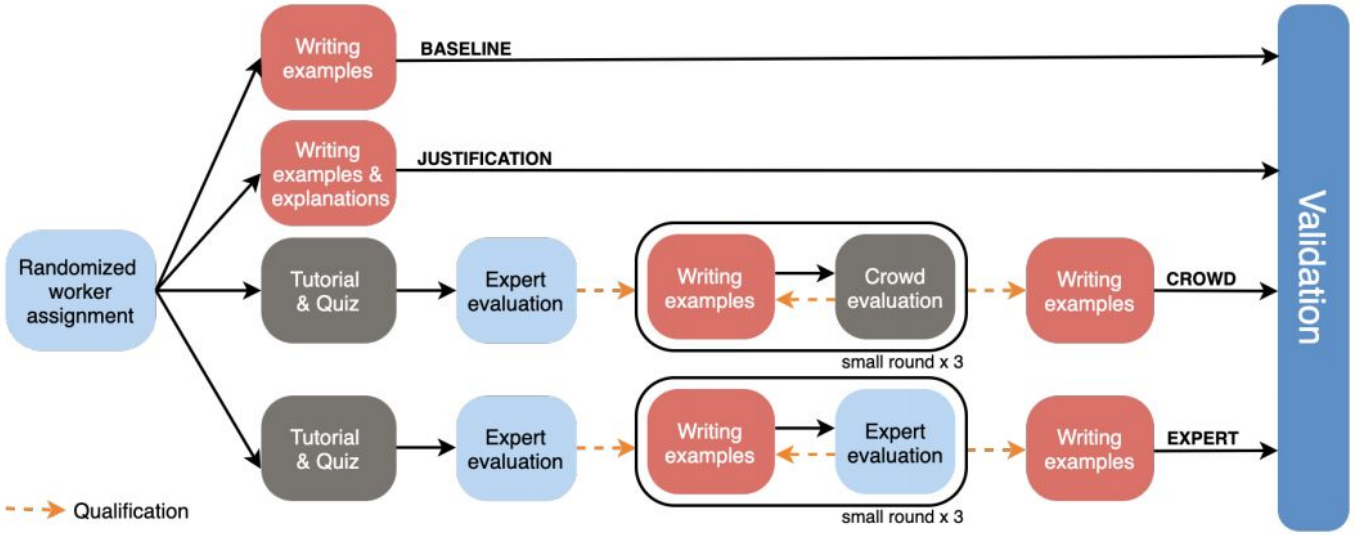


Figure 1: The initial pool of crowdworkers are randomly assigned to one of four protocols and the datasets are collected in parallel.

What Ingredients Make for an Effective Crowdsourcing Protocol for Difficult NLU Data Collection Tasks? (ACL 2021)

공정한 평가를 위한 Train/Dev/Test split

- **모델이 배워야하는 pattern을 제대로 학습해야 Test 성능이 오를 수 있도록 분리**
  - NER: 특정 단어와 label을 연관짓는 모델 X → 문맥 속에서 단어의 entity를 판단하는 모델 O
  - NLI: 가설 문장만 보고 label을 연관짓는 모델 X → 가설 문장과 전제 문장 모두를 보고 그 둘의 관계를 추론하는 모델 O
  - RE: 일반적인 상황에서는 no_relation 이 많기 때문에 전체 데이터에서 random sampling 하여 추출한 문장들로 test set 구성

공정한 평가를 위한 태스크 metric 선정

Language Conference

- **태스크와 데이터, 그리고 모델 특성의 측면을 다각도로 고려**
  - NER
    - entity-level macro F1: Tokenization 의 성능에 따라 점수가 크게 차이가 나는 metric
    - character-level macro F1: Tokenization 성능의 영향을 적게 받는 metric
  - MRC
    - EM: Tokenization 의 성능에 따라 점수가 크게 차이가 나는 metric
    - ROUGE-W:
      1) Tokenization 성능의 영향을 적게 받는 metric
      2) 우연히 character가 겹치는 단어를 정답으로 잡은 경우에 대해 과도하게 점수를 많이 주지 않는 metric
      (**한국**의 **위**인들 / **국한**된 범**위**)
  - DP
    - macro F1: class imbalance 가 심하기 때문

Baseline 코드 소개

LangCon 2021

Baseline Code

Language Conference

https://github.com/KLUE-benchmark/KLUE-baseline

KLUE Leaderboard

LangCon 2021

KLUE Leaderboard

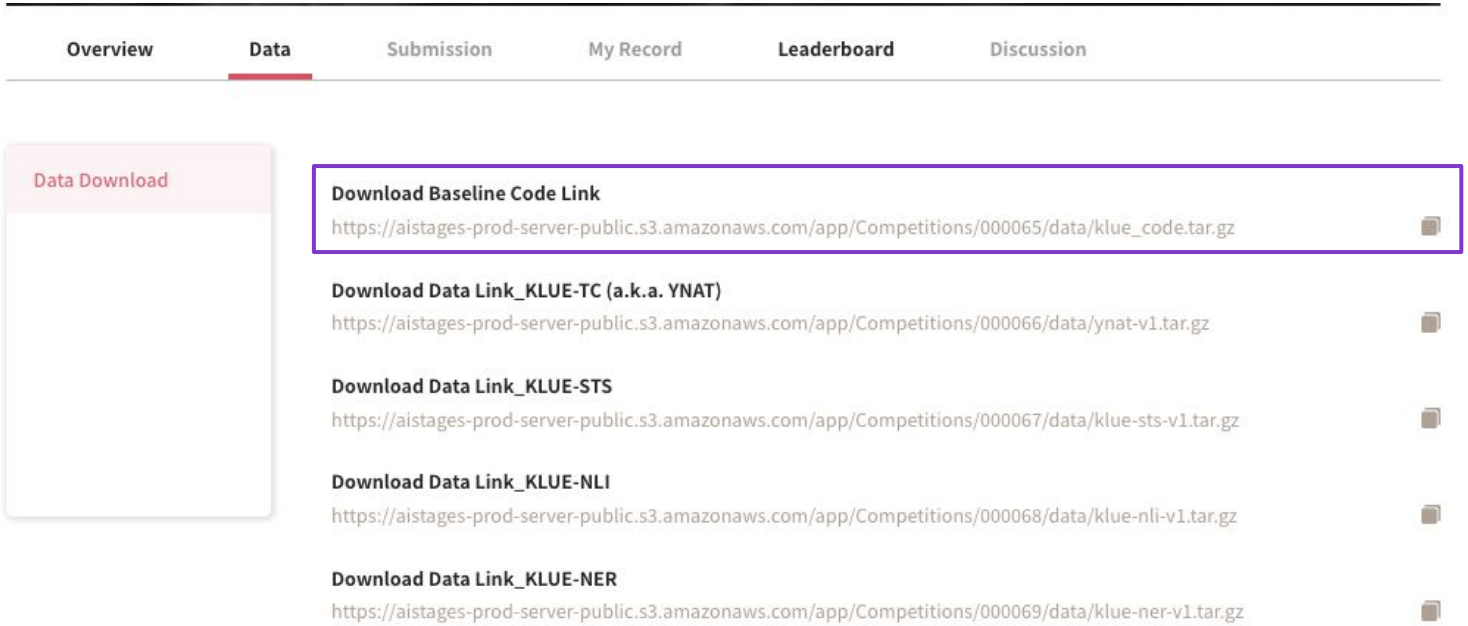
Language Conference

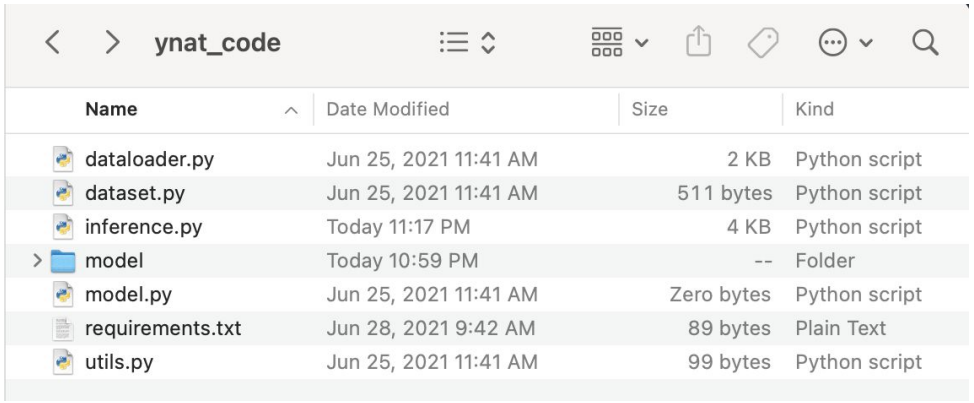
https://klue-benchmark.com/

YNAT 제출 예시

Language Conference

- Individual Tasks > YNAT > Data > Download Baseline Code > 수정 > 압축 > 제출

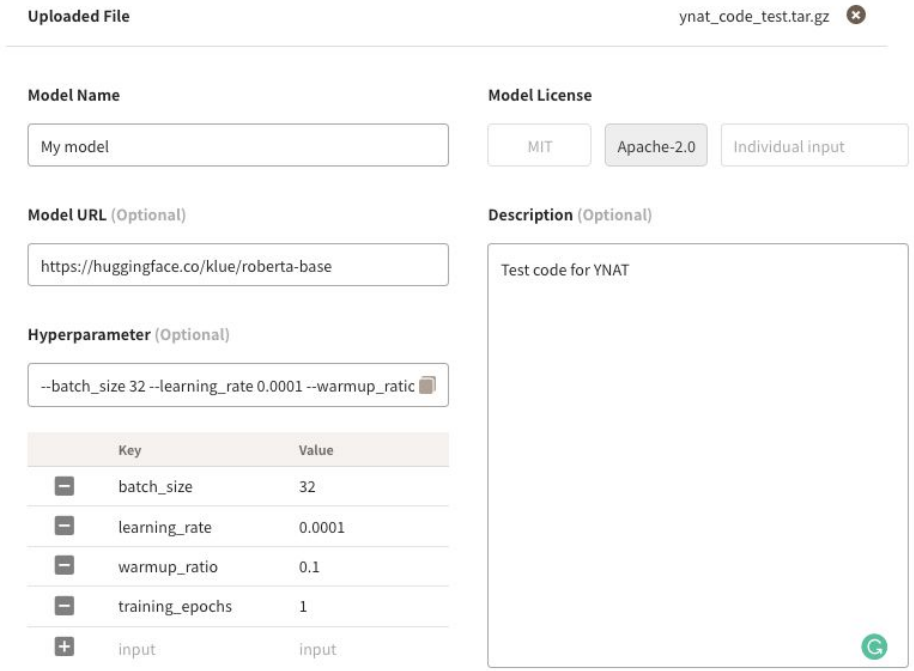


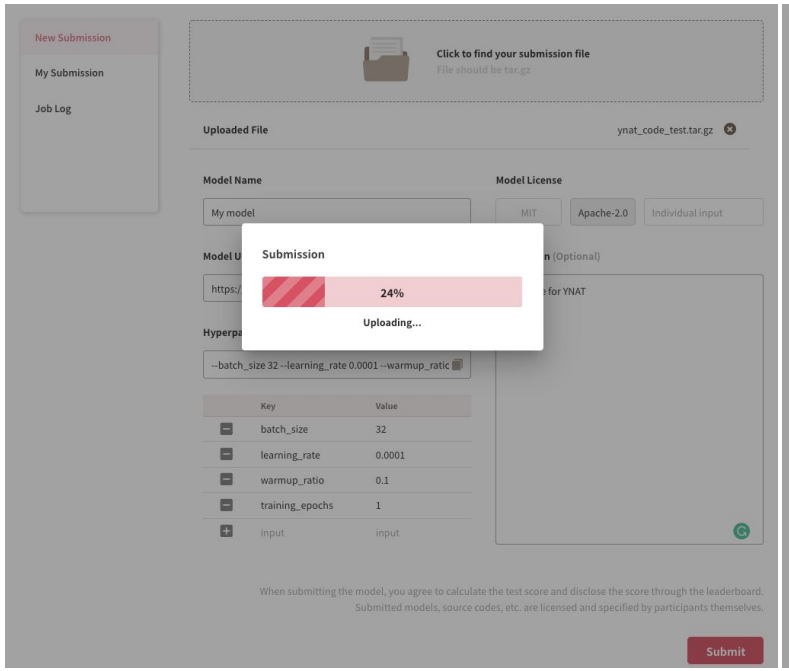


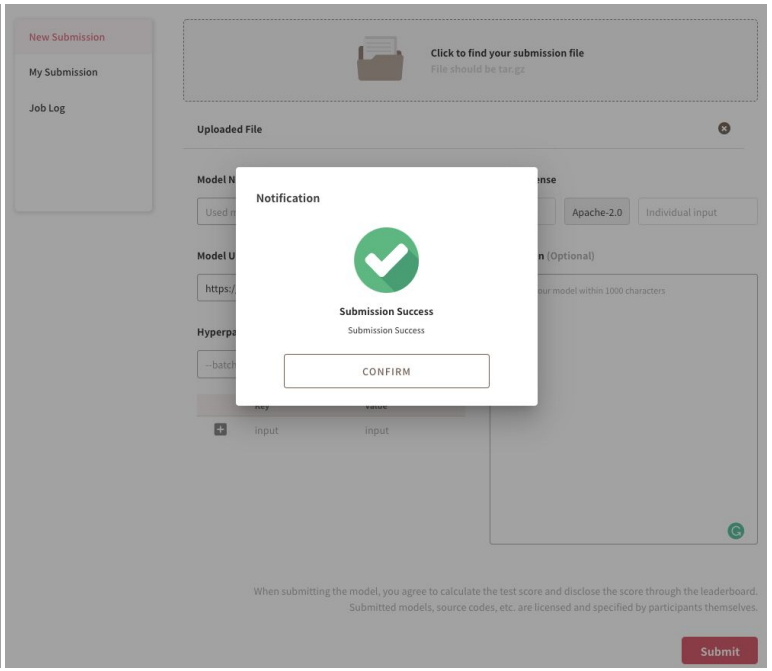
YNAT 제출 예시

Language Conference

- Individual Tasks > YNAT > Data > Download Baseline Code > 수정 > 압축 > 제출



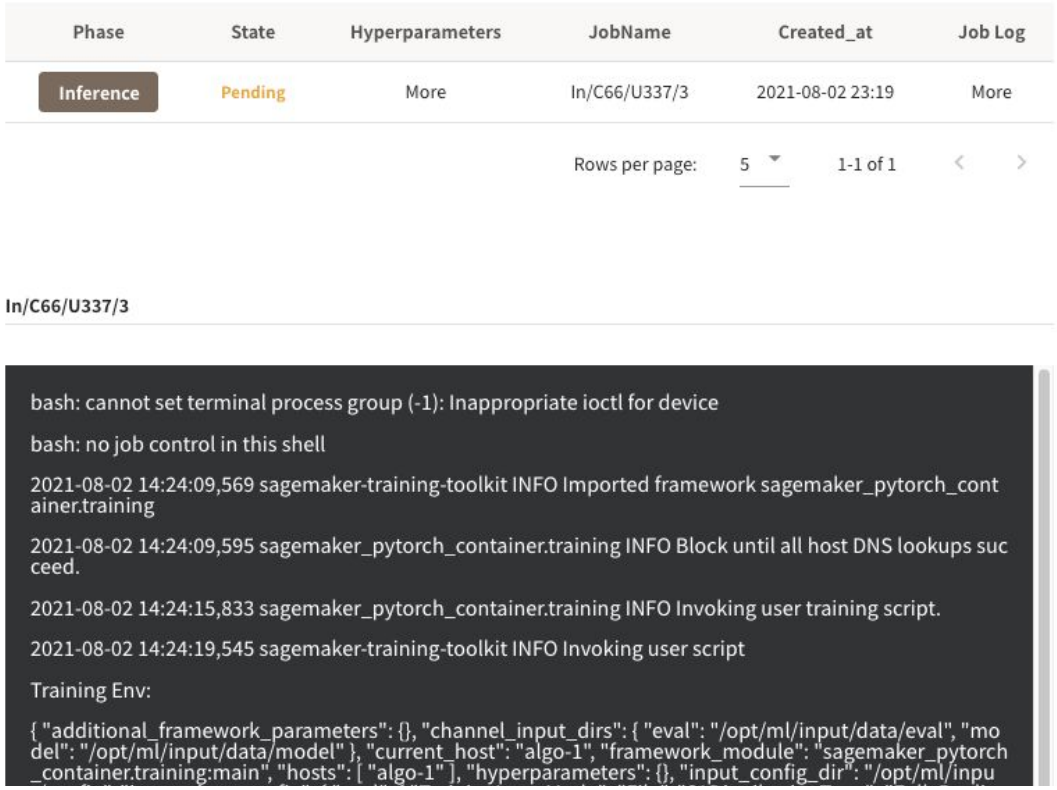


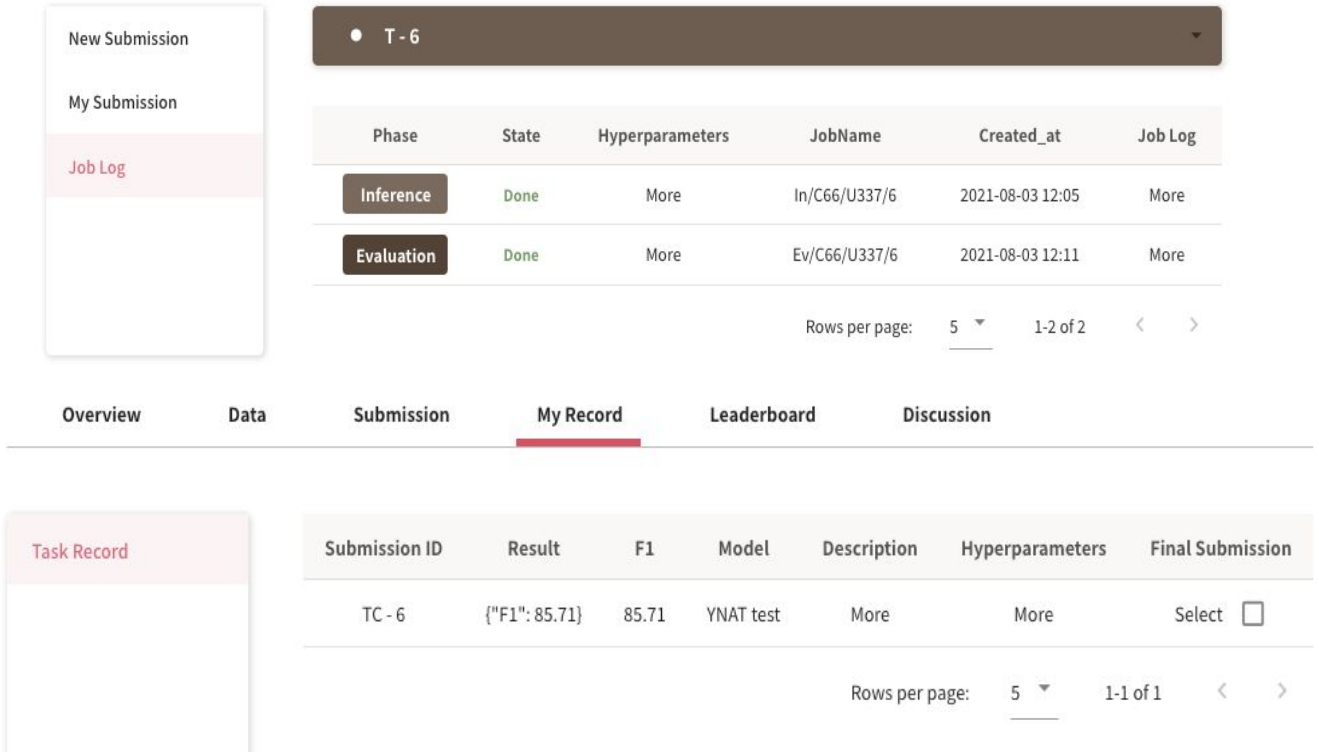


YNAT 제출 예시

Language Conference

- Individual Tasks > YNAT > Data > Download Baseline Code > 수정 > 압축 > 제출





KLUE 이후의 데이터(에 대한 개인적인 바람)

LangCon 2021

모델의 발전을 견인하는 데이터

- 데이터에 대한 모델 점수의 향상이 곧 모델의 성능 향상으로 이어지는 데이터
- Baseline 모델의 점수가 낮아서 점수 향상의 여유가 있는 데이터
- 문제에 대한 접근 방식에 대해 다르게 생각해보게 해주는 데이터
  - e.g., social bias inference
- static benchmark vs. dynamic benchmark

Dynabench: Rethinking Benchmarking in NLP

Douwe Kiela[†], Max Bartolo[‡], Yixin Nie[⋆], Divyansh Kaushik[§], Atticus Geiger[¶],

Zhengxuan Wu[¶], Bertie Vidgen[‖], Grusha Prasad[⋆⋆], Amanpreet Singh[†], Pratik Ringshia[†],

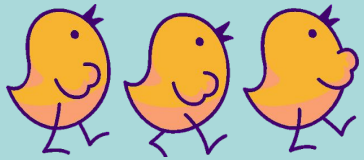
Zhiyi Ma[†], Tristan Thrush[†], Sebastian Riedel[†‡], Zeerak Waseem[††], Pontus Stenetorp[‡],

Robin Jia[†], Mohit Bansal[⋆], Christopher Potts[¶] and Adina Williams[†]

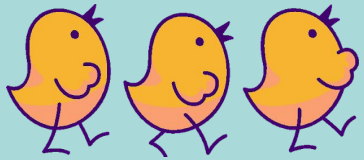
[†] Facebook AI Research; [‡] UCL; [⋆] UNC Chapel Hill; [§] CMU; [¶] Stanford University

[‖] Alan Turing Institute; [⋆⋆] JHU; [††] Simon Fraser University

dynabench@fb.com



Thank you!



Any questions?