Other Dataset

Jehyuk Lee

Department of AI, Big Data & Management

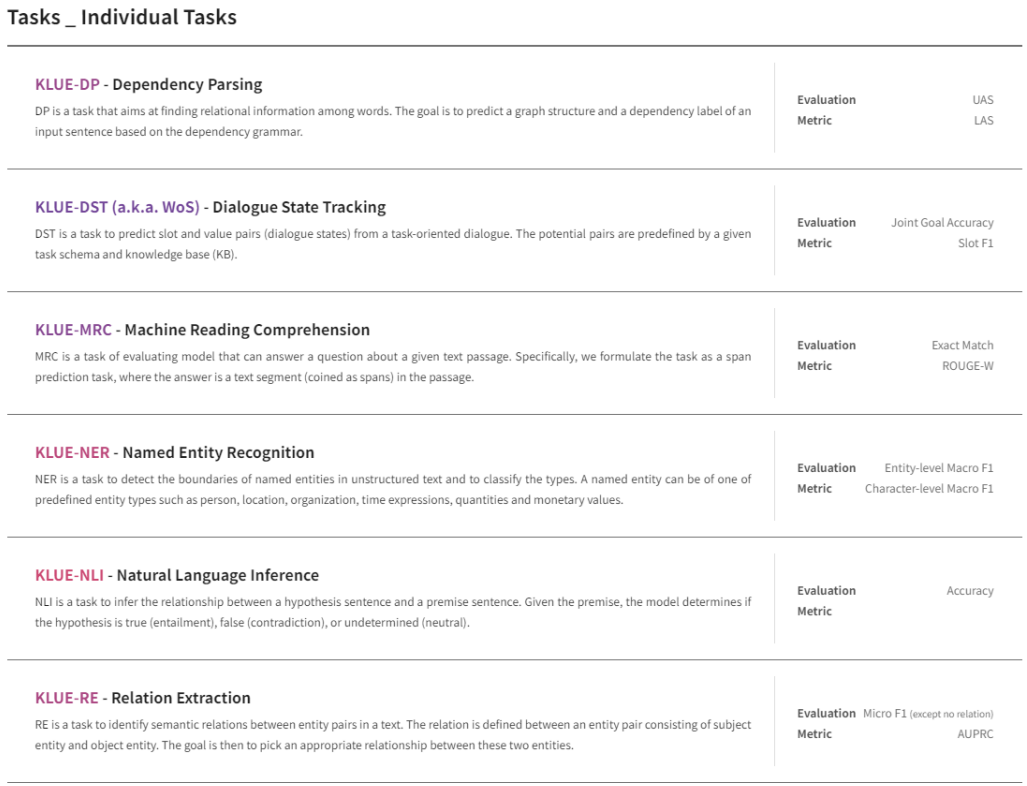
Kookmin University

Contents

- 이 Slide에서는 활용할 수 있는 데이터 셋을 몇가지 알려드리려 합니다.
  - 단, 한국어 데이터를 중심으로

KLUE

- 다양한 ML task에 대해서 학습 데이터로 사용 가능한 데이터
  - https://klue-benchmark.com/tasks 데이터도 있고 label 도 있으므로 사용하기 좋음
  - Topic Classification, Natural Language Understanding, Named Entity Recognition, …
  - GLUE의 한국어 버전 느낌…(GLUE도 찾아보세요)



Tasks _ Individual Tasks

KLUE-DP - Dependency Parsing
DP is a task that aims at finding relational information among words. The goal is to predict a graph structure and a dependency label of an input sentence based on the dependency grammar.
Evaluation: UAS
Metric: LAS

KLUE-DST (a.k.a. WoS) - Dialogue State Tracking
DST is a task to predict slot and value pairs (dialogue states) from a task-oriented dialogue. The potential pairs are predefined by a given task schema and knowledge base (KB).
Evaluation: Joint Goal Accuracy
Metric: Slot F1

KLUE-MRC - Machine Reading Comprehension
MRC is a task of evaluating model that can answer a question about a given text passage. Specifically, we formulate the task as a span prediction task, where the answer is a text segment (coined as spans) in the passage.
Evaluation: Exact Match
Metric: ROUGE-W

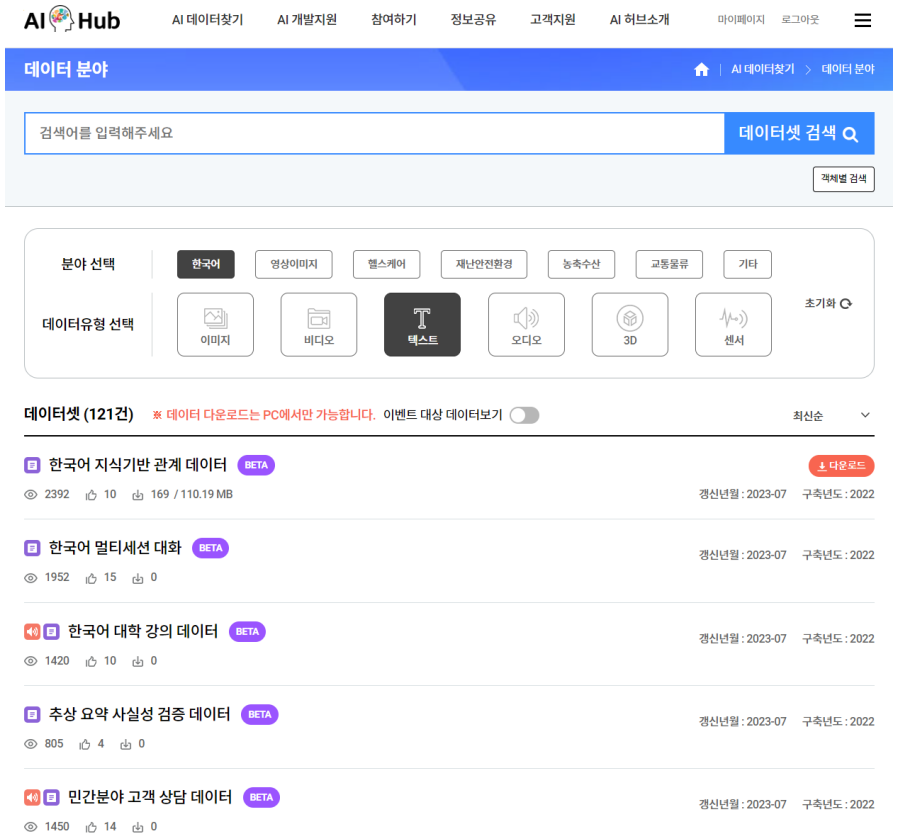
KLUE-NER - Named Entity Recognition
NER is a task to detect the boundaries of named entities in unstructured text and to classify the types. A named entity can be of one of predefined entity types such as person, location, organization, time expressions, quantities and monetary values.
Evaluation: Entity-level Macro F1
Metric: Character-level Macro F1

KLUE-NLI - Natural Language Inference
NLI is a task to infer the relationship between a hypothesis sentence and a premise sentence. Given the premise, the model determines if the hypothesis is true (entailment), false (contradiction), or undetermined (neutral).
Evaluation: Accuracy
Metric:

KLUE-RE - Relation Extraction
RE is a task to identify semantic relations between entity pairs in a text. The relation is defined between an entity pair consisting of subject entity and object entity. The goal is then to pick an appropriate relationship between these two entities.
Evaluation: Micro F1 (except no relation)
Metric: AUPRC

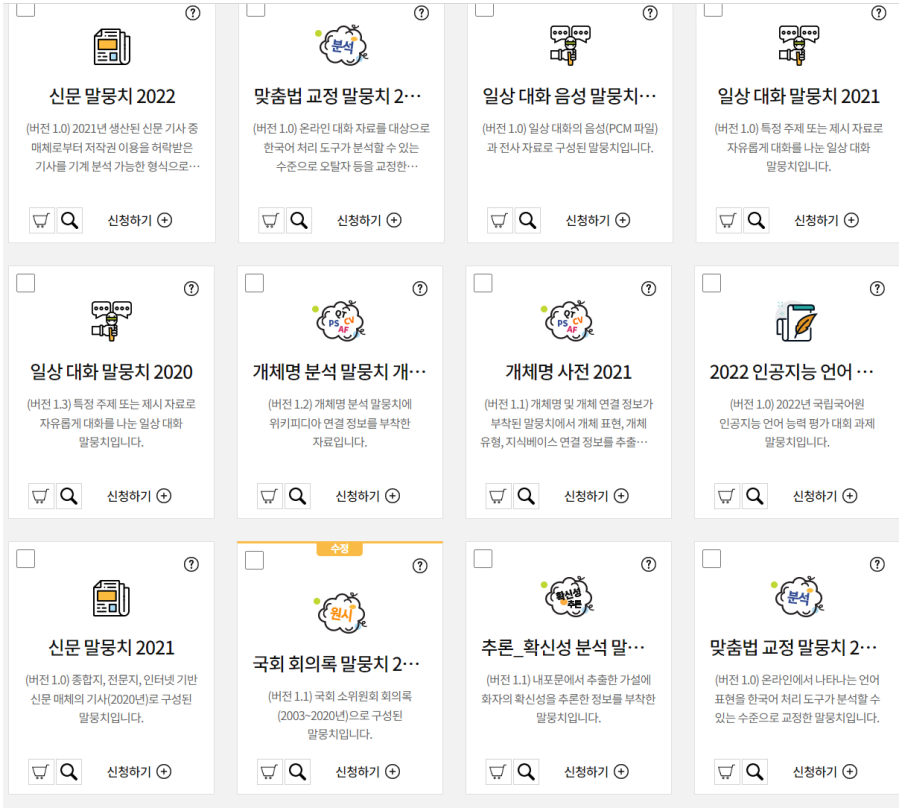
AI-HUB

- 다양한 task에 대해서 학습 데이터로 사용 가능한 데이터
  - 사용 방법은 각 데이터에 들어가서 별도로 보시면 됩니다.
    - beta버전 말고 정식 버전을 사용하세요.



모두의말뭉치

- 국립국어원에서 제공하는 말뭉치 데이터
  - 데이터 전처리, 사전 구축에 사용할 수 있음
    - 특히, 댓글 데이터를 사용한다면 구어체 전처리에 사용



모두의말뭉치

- 국립국어원에서 제공하는 말뭉치 데이터
  - 데이터 활용에 도움이 되는 List
    - https://corpus.korean.go.kr/resultRequest/supportDataList.do

전체 9건 | 10개씩 보기

번호	제목	작성자	작성 날짜	조회 수
9	모두의 말뭉치 동영상 강의 자료(마지막회, 어휘 의미 분석 말뭉치를 사용한 '먹다'의 출현 환경 탐색)	국립국어원	2022.11.21.	764
8	모두의 말뭉치 동영상 강의 자료(5회차, 일상 대화 말뭉치를 사용하여 '완전'의 부사적 용법 탐색)	국립국어원	2022.11.16.	1076
7	'모두의 말뭉치' 동영상 강의 자료(4회차, 분석 말뭉치 활용하기)	국립국어원	2022.11.04.	979
6	'모두의 말뭉치' 동영상 강의 자료(3회차, 원시말뭉치 활용하기)	시스템 관리	2022.10.28.	1176
5	'모두의 말뭉치' 동영상 강의 자료(2회차, 말뭉치 파일 탐색하기)	국립국어원	2022.10.21.	1386
4	'모두의 말뭉치' 동영상 강의 자료(1회차, 말뭉치 소개 및 파일 신청하기)	국립국어원	2022.09.27.	1226
3	'모두의 말뭉치' 한국어 빅데이터 활용 기업 특별 강연 발표 자료(눈으로 보는 음성 기록 클로바노트)	국립국어원	2022.09.22.	992
2	'모두의 말뭉치' 한국어 빅데이터 활용 기업 특별 강연 발표 자료(기계 번역 성능 향상을 위한 말뭉치 구축의 중요성)	국립국어원	2022.09.22.	625
1	'모두의 말뭉치' 한국어 빅데이터 활용 기업 특별 강연 발표 자료 (AI 챗봇 윤리 이슈와 대응 사례)	국립국어원	2022.09.22.	425

모두의말뭉치

- 국립국어원에서 제공하는 말뭉치 데이터
  - 데이터 활용 사례
    - https://corpus.korean.go.kr/resultRequest/useList.do

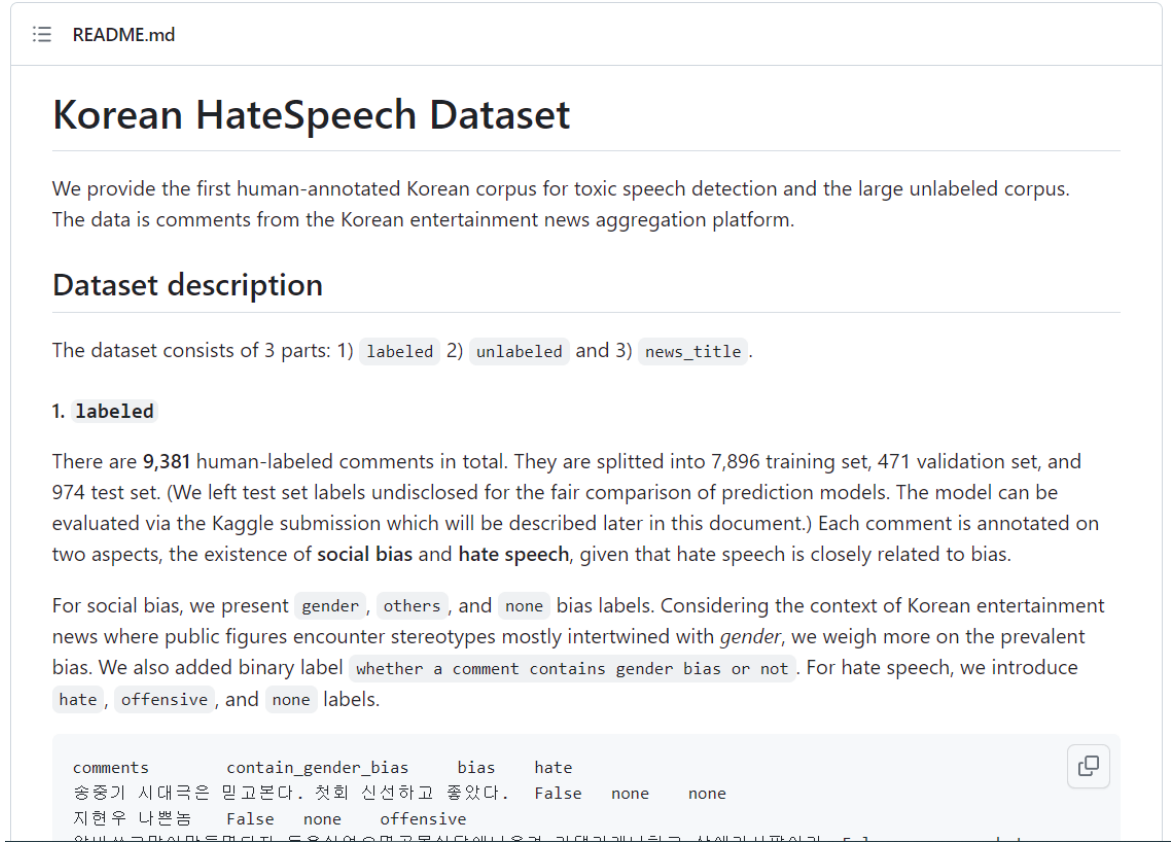
총 21건

10개씩 보기

번호	제목	결과물 형태	등록 날짜	조회 수
21	발화속도와 말차례 교체 빈도에 따른 운율 단위 변화에 관한 연구	학술 발표, 논문, 과제 보고서	2023.05.22.	117
20	Morpho-phonological effects on the phonetic characteristics of tense consonants in Korean compounds	학술 발표, 논문, 과제 보고서	2023.04.05.	69
19	PLM 기반 한국어 개체명 인식 (NER)	학술 발표, 논문, 과제 보고서	2023.03.07.	197
18	딥러닝 기반 한국어 개체명 인식의 평가와 오류 분석 연구	학술 발표, 논문, 과제 보고서	2023.03.07.	121
17	'-어 하-'의 통합 양상과 의미기능	학술 발표, 논문, 과제 보고서	2023.02.06.	108
16	사적 구어 환경에서 정도부사 '너무', '되게', '엄청'의 의미적 선호 분석 연구	학술 발표, 논문, 과제 보고서	2022.12.02.	135
15	비정형 텍스트 데이터에서의 개인정보 비식별화 대상 탐지	학술 발표, 논문, 과제 보고서	2022.11.23.	58
14	우왕좌왕 세종대왕	인공지능 모델/서비스/제품	2022.10.19.	341
13	DaramGPT	학술 발표, 논문, 과제 보고서	2022.09.14.	214
12	[석사학위논문] 연결어미와 조사 '은/는'의 결합에 대한 연구	학술 발표, 논문, 과제 보고서	2022.09.14.	152

Hate Speech Dataset

- Human-labeled Dataset



README.md

Korean HateSpeech Dataset

We provide the first human-annotated Korean corpus for toxic speech detection and the large unlabeled corpus. The data is comments from the Korean entertainment news aggregation platform.

Dataset description

The dataset consists of 3 parts: 1) `labeled` 2) `unlabeled` and 3) `news_title`.

1. `labeled`

There are **9,381** human-labeled comments in total. They are splitted into 7,896 training set, 471 validation set, and 974 test set. (We left test set labels undisclosed for the fair comparison of prediction models. The model can be evaluated via the Kaggle submission which will be described later in this document.) Each comment is annotated on two aspects, the existence of **social bias** and **hate speech**, given that hate speech is closely related to bias.

For social bias, we present `gender`, `others`, and `none` bias labels. Considering the context of Korean entertainment news where public figures encounter stereotypes mostly intertwined with *gender*, we weigh more on the prevalent bias. We also added binary label `whether a comment contains gender bias or not`. For hate speech, we introduce `hate`, `offensive`, and `none` labels.

```
comments        contain_gender_bias     bias    hate
송중기 시대극은 믿고본다. 첫회 신선하고 좋았다.   False   none    none
지현우 나쁜놈    False   none    offensive
```

End of the documents