NLP Benchmarks

Nitish Joshi, 21st February 2022

Outline

- Datasets in NLP, and useful resources to use them.
- Considerations when choosing a dataset.
- Challenges in data collection.

Individual Task Benchmarks

- Tasks: Machine Translation,Question Answering, Sentiment Analysis, Common Sense Reasoning, Summarization etc.

- http://nlpprogress.com - Useful resource to track datasets for different tasks in NLP

Individual Task Benchmarks

- What is different in all the benchmarks for the same task (say QA)?
  - Domain (e.g. sports domain vs legal domain)
  - Fine-grained phenomena (e.g. short answers vs long answers)
  - Language
  - Evaluation Metric (e.g. exact span match vs multiple-choice)
  - etc.

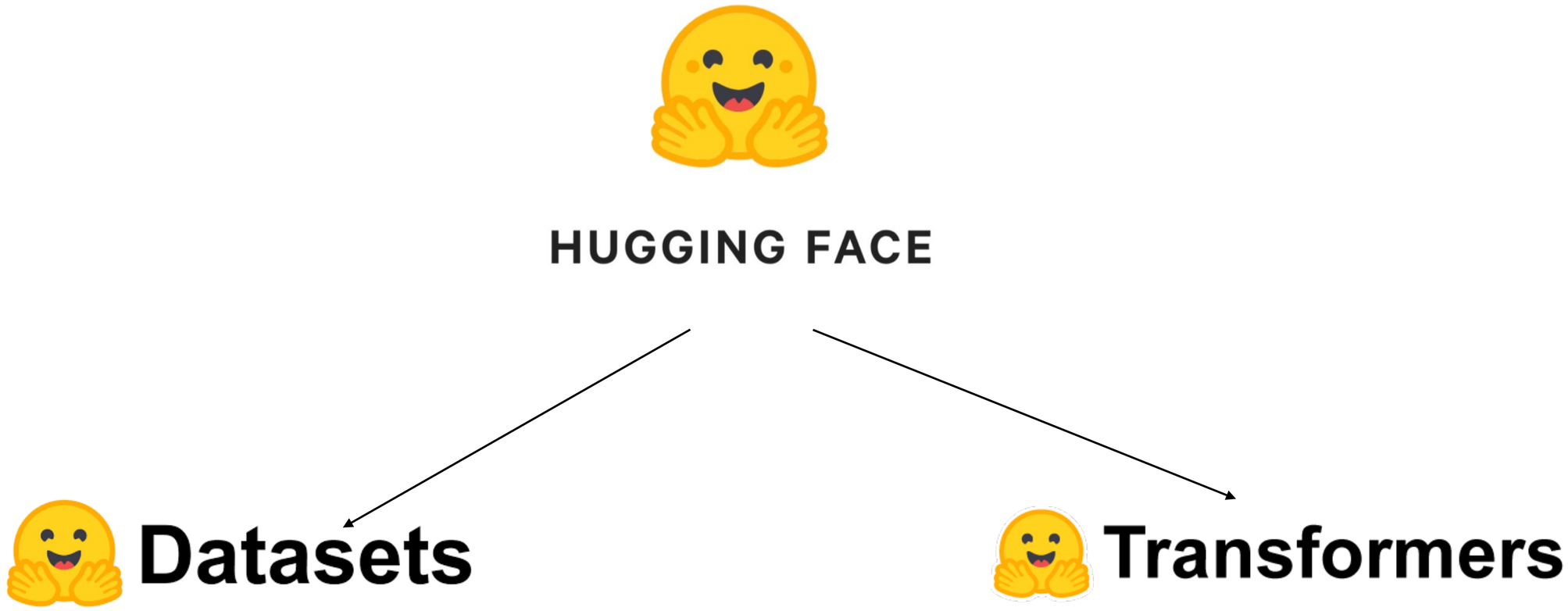
Multi-task Benchmarks

- GLUE (https://gluebenchmark.com) and SuperGLUE (https://super.gluebenchmark.com) include a suite a tasks designed to test natural language understanding
    - Tasks: Sentiment analysis, paraphrase detection, natural language inference etc.
- Highly influential in recent developments in NLP (BERT, GPT-2 etc) and developed at NYU!!

Multi-task Benchmarks

- BigBench (https://github.com/google/BIG-bench) — create a collaborative benchmark.
- Spans 204 diverse tasks including linguistics, common-sense reasoning, social bias, math etc.
- Influential in recent developments in large language models like GPT-3. (More later in the course!)

Useful Resources



Datasheets for Datasets

- Analogous to the datasheets common in electronic components (e.g. operating characteristics, usage etc.)
- Why? Increases transparency and accountability.
- Gebru et al., 2019 : Standardizes dataset documentation along:
    - Creation
    - Composition
    - Intended uses
    - Maintenance

Gebru et al., 2019

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should *not* be used?

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

How many instances of each type are there?

Gebru et al., 2019

Dataset Distribution

How is the dataset distributed? (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?)

When will the dataset be released/first distributed? (Is there a canonical paper/reference for this dataset?)

Data Preprocessing

What preprocessing/cleaning was done? (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.)

Was the "raw" data saved in addition to the preprocessed/cleaned data? (e.g., to support unanticipated future uses)

Legal & Ethical Considerations

If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.)

If it relates to other ethically protected subjects, have appropriate obligations been met? (e.g., medical data might include information collected from animals)

If it relates to people, were there any ethical review applications/reviews/approvals? (e.g. Institutional Review Board applications)

Gebru et al., 2019

Challenges in Dataset Construction

- Annotation Artifacts in Datasets (Gururangan et al., 2018)
- Annotators might use simple rules or heuristics to create the examples
- Task: Given a premise p write three hypothesis h such that:

Entailment	h is definitely true given p
Neutral	h might be true given p
Contradiction	h is definitely **not** true given p

Challenges in Dataset Construction

Contradiction:

Premise: The woman was standing near the shop.
Hypothesis: The woman was not near the shop.

Premise: She is selling bamboo sticks.
Hypothesis: She is not taking money for the bamboo sticks.

Premise: It was raining heavily today.
Hypothesis: There was no water on the ground today.

Notice anything common?

Challenges in Dataset Construction

Contradiction:

Premise: The woman was standing near the shop.
Hypothesis: The woman was **not** near the shop.

Premise: She is selling bamboo sticks.
Hypothesis: She is **not** taking money for the bamboo sticks.

Premise: It was raining heavily today.
Hypothesis: There was **no** water on the ground today.

Annotators tend to add negation words in contradiction

Challenges in Dataset Construction

Contradiction:

Premise: The woman was standing near the shop.
Hypothesis: The woman was **not** near the shop.

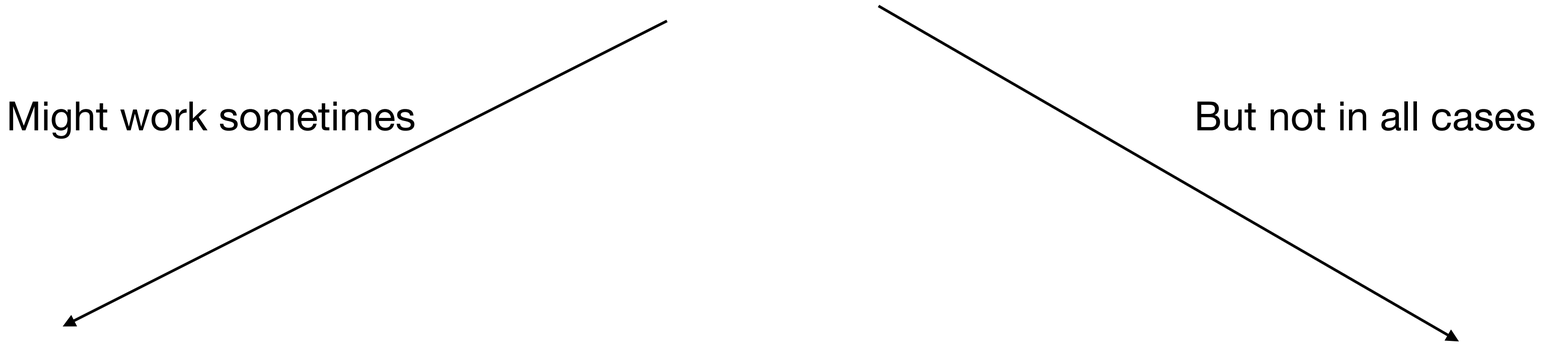
Premise: She is selling bamboo sticks.
Hypothesis: She is **not** taking money for the bamboo sticks.

Premise: It was raining heavily today.
Hypothesis: There was **no** water on the ground today.

- Models trained on this data may predict contradiction whenever negation word is present.
- Why might this be bad?

Challenges in Data Collection

Heuristic	Definition
Negation Word	Assume that the label is contradiction whenever a negation word is present in the hypothesis.



Premise: The woman was standing near the shop.
Hypothesis: The woman was not near the shop.
Label: Contradiction

Premise: The actor paid by the doctor.
Hypothesis: The doctor did not treat the actor.
Label: Neutral

Challenges in Data Collection

Heuristic	Definition
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise

Might work sometimes

But not in all cases

Premise: The woman was standing near the shop.
Hypothesis: The woman was near the shop.
Label: Entailment

Premise: The doctor was paid by the actor.
Hypothesis: The doctor paid the actor.
Label: Not Entailment

Spurious Correlations in Datasets

- Certain input features (e.g. negation words) are highly correlated with a certain label (e.g. contradiction).
- Is my model right the right reasons? (McCoy et al., 2019)
- If the model relies on the spurious correlations, then it may not generalize well when used in practice!

Summary

- Single-task vs Multi-task benchmarks
- Huggingface Datasets Library
- Datasheets for Datasets
- Challenges in data collection — annotator artifacts.