HSLU Lucerne University of Applied Sciences and Arts

Course Project

NLP
Andreas Marfurt

Information Technology
20.02.2025

FH Zentralschweiz

Goals

- Use the models right after you've learned about them
  - Projects run during the course
- Gain model training expertise
  - Single-person projects
- Hands-on experience with a real benchmark
  - What are the specialties of each model type?
  - How well do the models perform on current benchmarks?
- Structure experiments to not lose track
  - Which hyperparameters, code version, model was used?
  - Using a tool (in this course: Weights and Biases) makes this easier
- Document results so other people can understand and reproduce them
  - Documentation of code and training (Jupyter notebooks and W&B)
  - Present and discuss the important points (presentation)

Task: Commonsense QA

- Commonsense question answering
- Question and 5 answer options
- Paper: Talmor et al., 2019
- Part of LM evaluation harness

Examples

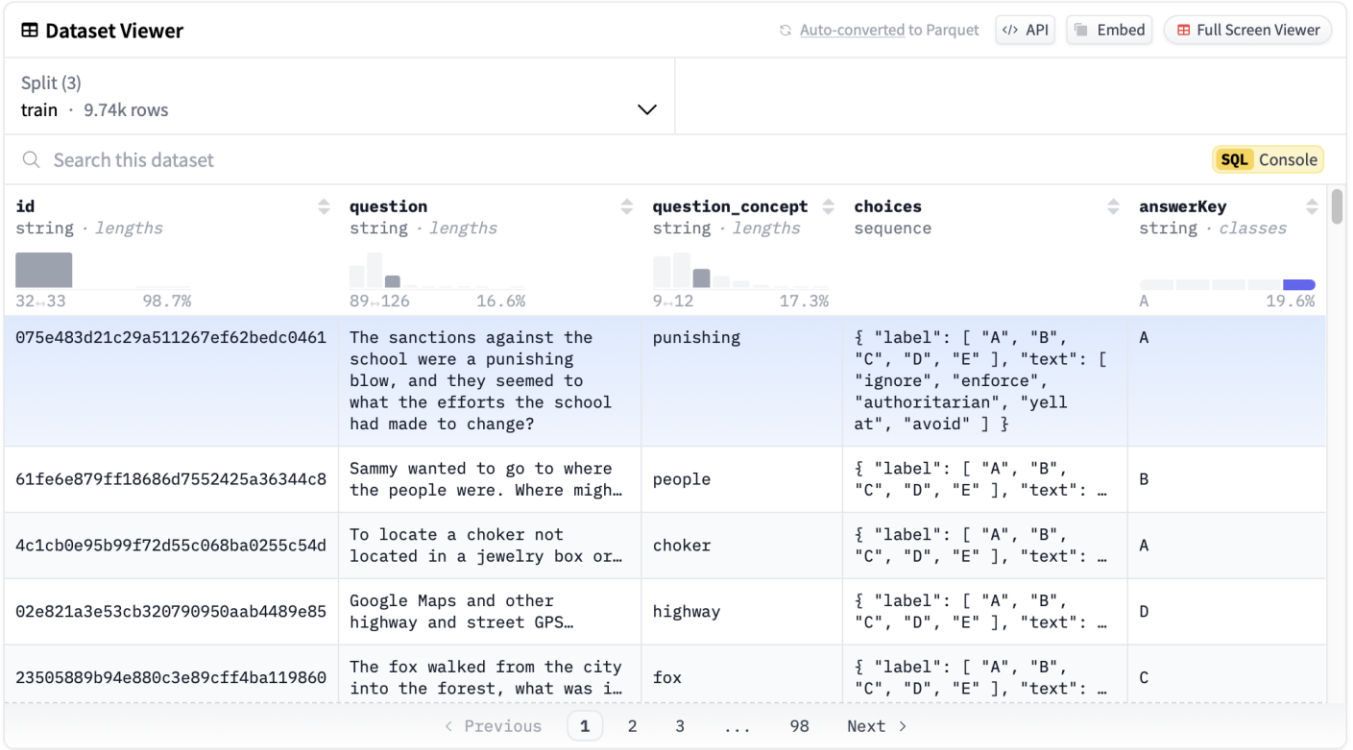
- Question: The sanctions against the school were a punishing blow, and they seemed to what the efforts the school had made to change?
- Choices
  - A: ignore
  - B: enforce
  - C: authoritarian
  - D: yell at
  - E: avoid
- Answer: A

Examples

- Question: What home entertainment equipment requires cable?
- Choices
  - A: radio shack
  - B: substation
  - C: cabinet
  - D: television
  - E: desk
- Answer: D

Data

- Available on Hugging Face: https://huggingface.co/datasets/tau/commonsense_qa
- Only train and validation splits have an answerKey
- We will use our own dataset splits



Required Capabilities

Category	Definition	%
Spatial	Concept A appears near Concept B	41
Cause & Effect	Concept A causes Concept B	23
Has parts	Concept A contains Concept B as one of its parts	23
Is member of	Concept A belongs to the larger class of Concept B	17
Purpose	Concept A is the purpose of Concept B	18
Social	It is a social convention that Concept A correlates with Concept B	15
Activity	Concept A is an activity performed in the context of Concept B	8
Definition	Concept A is a definition of Concept B	6
Preconditions	Concept A must hold true in order for Concept B to take place	3

Table 3: Skills and their frequency in the sampled data. As each example can be annotated with multiple skills, the total frequency does not sum to 100%.

Question Words

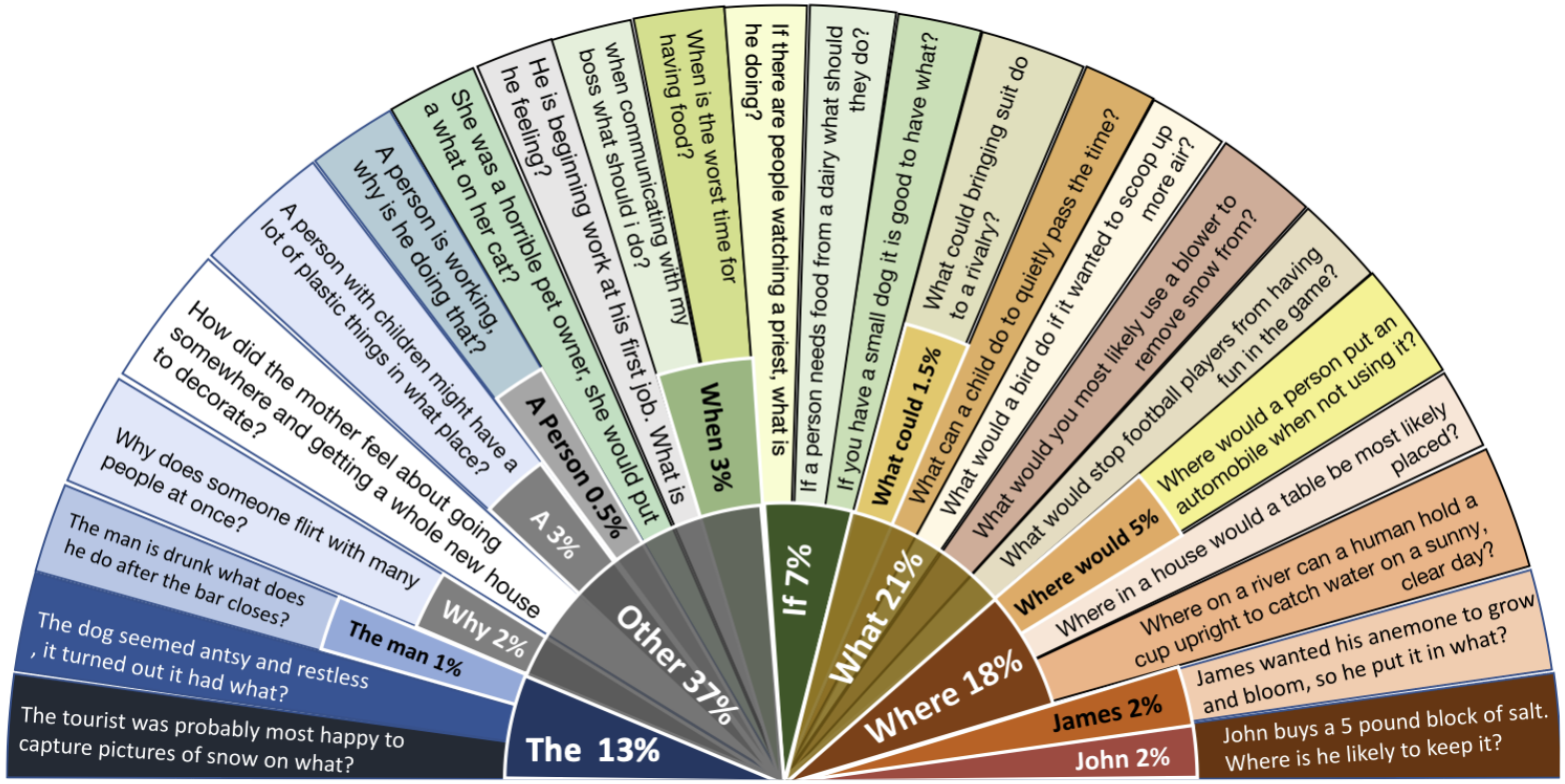


Figure 4: Distribution of the first and second words in questions. The inner part displays words and their frequency and the outer part provides example questions.

https://aclanthology.org/N19-1421.pdf

Train/Validation/Test Splits

```
>>> from datasets import load_dataset

>>> train = load_dataset("tau/commonsense_qa", split="train[:-1000]")
>>> valid = load_dataset("tau/commonsense_qa", split="train[-1000:]")
>>> test = load_dataset("tau/commonsense_qa", split="validation")

>>> print(len(train), len(valid), len(test))
8741 1000 1221
```

Train/Validation/Test Splits

- Train: Use this data to train your models
- Validation: Compare different models/model settings
  - Hyperparameter tuning/optimization
  - This data is not used to train your model
  - Select the best model and hyperparameters
- Test: Test it on the test data for the final comparison
  - Do not tune hyperparameters on the test set!
    - Considered cheating (in industry/research)
    - Will make your model look better than it will perform on truly unseen data

Experiment Tracking

- Make sure you invited me and the 4 TAs to your W&B team
- Create a view or report for each project
- Add the link to the introduction section of your Jupyter notebook

Available Infrastructure

Free, 1 GPU per user:

- HSLU GPUHub: https://gpuhub.labservices.ch/
  - Log in every 24h so that the virtual machine doesn't stop
  - Maximum runtime of 72h
  - Need to be in HSLU network (use VPN from outside)
  - Documentation
- Google Colab: https://colab.research.google.com/
  - Can be killed at any time by Google
  - Save your work regularly (e.g. in your Google Drive)

Deliverables and Deadline

- Deliverables
  - Self-contained Jupyter notebook
    - Includes experiment tracking link
  - Presentation slides as PDF
- Submissions on Ilias
  - Project 1: Week 8
  - Project 2: Week 14
- Presentations according to schedule (will be uploaded before deadline)

Including Prior Work

- It is ok to use prior work (code & models), but it needs to be cited!!!
  - This includes Hugging Face pretrained/finetuned models, Kaggle notebooks, code from blog posts, …
  - Clearly separate your own work from work done by others
  - Not citing is plagiarism, leads to failing course
    - In severe or repeat cases being expelled from the university
- Using a code completion tool is fine
  - You need to understand what your code is doing
    - You are responsible for your code, not the tool you used
  - State all tools used at the end of your notebook

Grading

- Jupyter notebook:
  - Notebook: see grading checklist
  - Project 1 “things to improve” implemented
  - Experiment tracking
  - Submission on time
    - Otherwise grade deduction (only acceptable reason for an extension is a doctor's note)
  - Form: Be concise but complete

- Presentation & discussion:
  - All important decisions presented
    - We will learn what is important
  - Clear explanations
    - Showing code usually not necessary
  - Correctness
  - Faithful to own work
  - Reflection/ideas for improvements

  - Good answers to questions
  - Submission on time
  - Form: Stay within time limit

Not graded: Model's accuracy