Benchmarks

Measuring the performace of LMs

Judging the quality of a LM

- Central problem: LMs only ‘see’ word forms, not meanings
- To which extent do LMs represent word meaning? Do they **understand**?
- See e.g. Bender and Koller (2020), who quote Devlin et al. (2019):

  “In order to train a model that **understands** sentence relationships, we pre-train for a binarized next sentence prediction task.”

- Or even worse, the media:

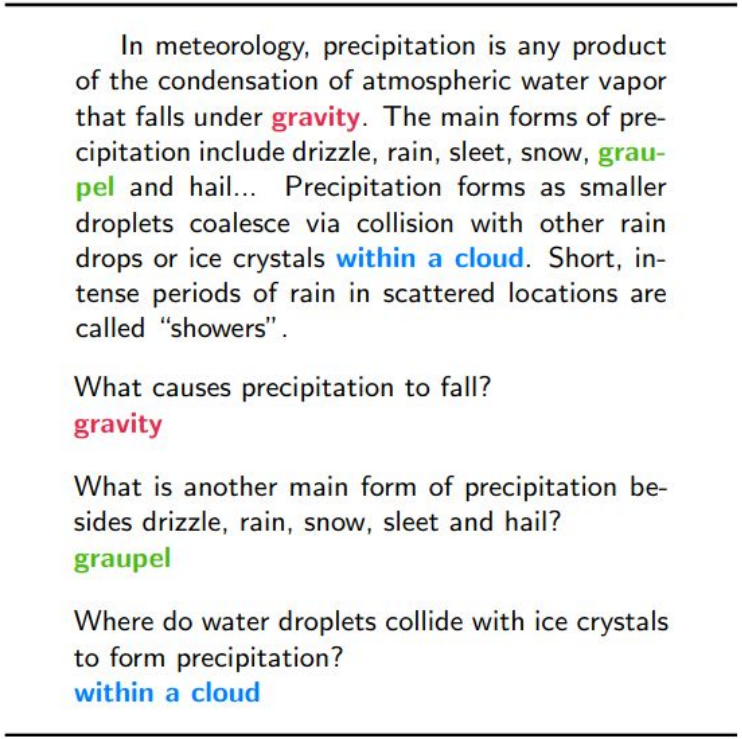
  “BERT is a system by which Google’s algorithm uses pattern recognition to better understand how human beings communicate so that it can return more relevant results for users.” (https://www.business2community.com/seo/what-to-do-about-bert-googles-recent-local-algorithm-update-02259261)

Task collections as benchmarks

- Practical perspective: Find tasks that focus on understanding, test model performance on these tasks
- Example benchmarks:
  - GLUE (Wang et al. 2019)
  - SuperGLUE (Wang et al. 2020)
- Result: Numerical score as performance measure
- Caveat: What does very high performance mean? Humans produce no perfect results either

Example tasks from GLUE

- Question answering
- SQuAD dataset: Given a text and a question, which part of the text answers the question?



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

Where do water droplets collide with ice crystals to form precipitation?
within a cloud

Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

IMG: Rajpurkar et al. 2016 (https://arxiv.org/pdf/1606.05250.pdf)

Example tasks from GLUE

- sentiment classification (Does a sentence have positive or negative sentiment?)
- sentence acceptability (is a sentence grammatical?)
- sentence similarity (How similar are two sentences?)

Goodhart’s Law

“When a measure becomes a target, it ceases to be a good measure." (Strathern 1997)

- applicable across domains
- Why is this a problem for language model evaluation?

Goodhart’s Law

“When a measure becomes a target, it ceases to be a good measure." (Strathern 1997)

- applicable across domains
- Why is this a problem for LM evaluation?
- Indirectly, LMs could be optimized towards LM objectives (MLM, NWP, NSP) **and** benchmark performance
- Just because an LM performs well on a benchmark, performance on other tasks is not guaranteed

Fine-Tuning vs. X-shot learning

- fine-tuning updates the model weights, specialization on particular task)
- few-shot, one-shot, zero-shot learning relies on forward-passes

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée
4  plush girafe => girafe peluche
5  cheese =>                           ← prompt
```

IMG source: Brown et al. (2020)

Practical Session

- Visit https://gluebenchmark.com/ or https://super.gluebenchmark.com and get familiar with the websites.
- Take a look at the datasets. Is your intuition about individual examples in line with the gold answers? What do you think about the data quality?