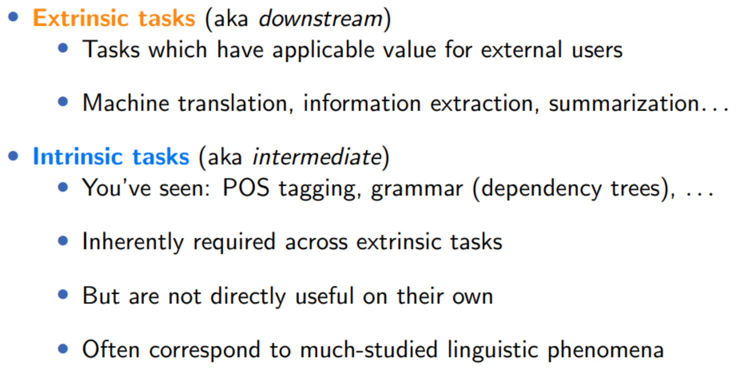
Advanced NLP Exercise 1

Yonatan

Open Questions

1. QA

We've seen the definitions of intrinsic and extrinsic tasks in the first lecture:



- Extrinsic tasks (aka *downstream*)
  - Tasks which have applicable value for external users
  - Machine translation, information extraction, summarization. . .
- Intrinsic tasks (aka *intermediate*)
  - You've seen: POS tagging, grammar (dependency trees), . . .
  - Inherently required across extrinsic tasks
  - But are not directly useful on their own
  - Often correspond to much-studied linguistic phenomena

Three examples of QA datasets that use QA to annotate concepts:

1. Google's boolq - link
   Answers for yes/no questions, triplets of (question, passage, answer), tests whether a model is NLI-capable, deep understanding of passage for answer.

2. Stanford Question Answering Dataset (SQuAD) - link
   Focuses on reading comprehension, like boolq there's no extrinsic task goal.

3. NarrativeQA - link
   Again, reading comprehension – specifically on (long) stories. Being able to answer questions in a very long text.

2. Inference-time scaling

a. We've talked about a few Inference-time Scaling methods in the third lecture:

   1. Self-consistency
      - Samples a diverse set of paths and answers the most consistent one, by "majority vote".
      - The main advantage is that accuracy becomes a lot better.
      - A computational bottleneck is increasing test-time computing (running on paths means more computing, meaning more time computing).

- Can be parallelized, as each path is different, and isn't dependent on others.

2. Verifiers
    - Verifying the validity of the answer, whether by RegEx, tests or other models entirely. We've seen it is used well with Self-consistency (selecting the best of verified answers, instead of all generated answers).
    - The main advantages are:
      1). Getting more valid answers (user-expectation and accuracy-wise!)
      2). We've seen in class automatic verifiers are possible to be trained and be used at test time – meaning better efficiency.
    - Computational bottlenecks:
      1). Using verifiers means there exists another layer of computation on the outputs, meaning increase in test-time computation.
      2). The time of computation is relative to the algorithm used, which may change for each "input", like calling a regex check, or an entire model.
    - Can technically be parallelized, as it is possible to verify a bunch of outputs in parallel – but the verification considers all the options, meaning it **cannot** be parallelized (waits for all generations).

3. Smaller models
    - Using n smaller models, that surpass the capabilities of a large model (n outputs vs. one).
    - The main advantages are:
      1). The same computing power, for more outputs
      2). Often better output (bound by verifiers quality)
    - Computational bottlenecks:
      1). How good the method is, is bound by how good the verifiers are – and so the computation times.
      2). Compute might be larger, the more small models are used.
      3). Affected by the number and quality of smaller models
    - Yes, it is parallelizable – as getting outputs by different models isn't influenced by each other.

4. O1 and R1 Models (R1 ~probably is an open-source replication of O1)
    - Using planning, backtracking and self-evaluation.
    - The main advantages are:
      1). Recognizes mistakes
      2). breaks down steps to simpler ones
      3). changes approaches when not working
      4). makes the model better at reasoning

- The main computational bottleneck is as we've seen in class: the output becomes larger as time passes, meaning more time and more compute is needed.
- Unfortunately, parallelization is not an option, as the model's behavior changes on its own output; each token may change its output, and the different behavior.

b. I would choose the Self-consistency method in that situation, as using a single GPU means I need parallelization – with no problem of memory, as I have large memory capacity.
   As I've mentioned before – this method allows for great rise of accuracy, can be parallelized and even be used with verifiers – that allow the output to be better, and overall be more efficient.

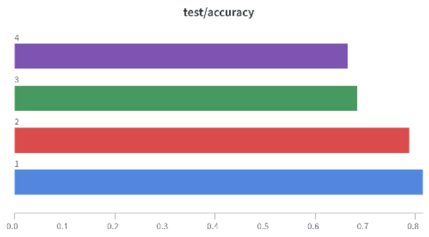
Programming Exercise

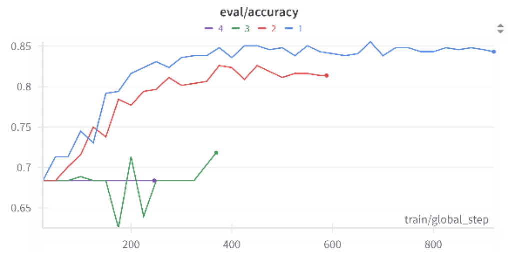
1. Link to (public) git:

https://github.com/YonatanGH/ANLP-EX1

2. Qualitative analysis

Here are the W&B charts of evaluation accuracy, and test accuracy:





We can see that indeed, the configuration that achieved the best validation accuracy also achieved the best accuracy (notice the names agree with the colors of the given train_loss.png)

As expected, the best run is better at determining which sentences DO NOT correlate. I can determine that, as most irregularities are where the best classifies 0, and the worst classifies 1.

But in context of finding what causes the irregularities, I have taken 4 examples of each possible type:

Let’s look at 1 (the best run), and 4 (the worst run), and try to deduce what type of examples were harder for 4:

Sentence 1	Sentence 2	Best run classification	Worst run classification
If the magazine lost more than $ 4.2 million in a fiscal year , O 'Donnell would be allowed to quit .	If Rosie lost more than $ 4.2 million in a fiscal year , O 'Donnell - by contract - would have been permitted to quit .	1	1
Shares of LendingTree rose 22 cents to $ 14.69 and have risen 14 percent this year .	Shares of LendingTree rose $ 6.03 , or 41 percent , to close at $ 20.72 on the Nasdaq stock market yesterday .	0	0
In his speech , Cheney praised Barbour 's accomplishments as chairman of the Republican National Committee .	Cheney returned Barbour 's favorable introduction by touting Barbour 's work as chair of the Republican National Committee .	1	0
Hong Kong was flat , Australia , Singapore and South Korea lost 0.2-0.4 percent .	Australia was flat , Singapore was down 0.3 percent by midday and South Korea added 0.2 percent .	0	1

It seems like the worse run has a problem of understanding location (no meaning of location), as in the fourth example, it mistook Australia and Hong Kong.

Also, It might take heavy weight on the last part of a sentence with the possibility of score of general similariy of a sentence, like seen in all examples.

In general, it seems the worse model would have a problem with sentences that contain similar words, with either very small changes in the sentence’s build, or changes in its beginning or ending although the same.