ANLP ex1

itamar.miron

May 2025

1 Open Questions

1.1

(a) **property: Recognizing text entailment (NLI)**
Dataset:ANLI (Adversarial Natural Language Inference) This setup evaluates a model's ability to determine whether a given hypothesis logically follows (entailment), contradicts (contradiction), or is neutral with respect to the premise. It tests a model's understanding of factual relationships and language reasoning rather than surface-level keyword matching.

(b) **property: coreference**
Dataset: DROP this dataset requires resolving pronouns and noun phrases over long passages to answer questions correctly, often involving reasoning over multiple references to the same entity—coreference resolution is key.

(c) **property: POS (Part-of-Speech) tagging**
Dataset: WorldTree includes questions that require syntactic disambiguation, where understanding the part of speech of a word (like noun vs verb) is essential to interpret the sentence meaning and answer correctly.

1.2

(a)
- **Chain-of-thought description:** "ask" the model to share the chain of thoughts on the answer
  advantages: improve results and don;t require extra model's run.
  computational bottlenecks: the output (and in most cases the input) will probably be longer wich will result in a proportional longer expected run time.
  can be parallelized? no, since chain of thoughts is by definition a chronological routine.
- **Self-consistency description:** Run the model a few times and take the majority the best answer.

advantages: in numerical tasks this can provide a more reliable results.
computational bottlenecks: the multiple run of the model will consume more compute and perhaps the selection of the ”best” will demand a classifier that will take it’s own compute resources.
can be parallelized? the parallelization of this method in the multiple model runs is depend on the number of available GPU’s, the post process can only occur after the model’s multiple runs completion.

- **Least-to-most description:** instead of asking the model to solve a problem, ask the model in the iterative way to only solve the next step of the problem and say what the next step sould be and than preform it.
  advantages: improve results and draw a path of the solution.
  computational bottlenecks: memory limitation of the model is limiting complex prompts from executing this method in some scenarios.
  can be parallelized? no, in the same manner like cot chronology is a key.
- **Self-ask description:** ask the model in the end of the prompts if any follow up questions needed.
  advantages: no constrains on the models answer structure. **computational bottlenecks:** if a followup question is needed, there is need for extra runs of the model in a nonparallel manner.
  can be parallelized? as mentioned above, no.

(b) In such scenario I will be using the Least to most method since it can’t be parallelized which compile well with the sole GPU case and it’s main computational bottleneck is memory-related for which the Large GPU memory capacity can handle it well.