Advanced NLP Exercise 1

Dean Tahory

1 Open Questions

Question 1

Quoref (Coreference Resolution)

- **Link:** `https://huggingface.co/datasets/allenai/quoref`
- **Explanation:** Quoref evaluates the intrinsic ability to resolve coreferences by requiring models to track references between entities in a passage, essential for coherent understanding of text.

QA2D (NLI)

- **Link:** `https://huggingface.co/datasets/domenicrosati/QA2D`
- **Explanation:** By transforming QA pairs into premise–hypothesis examples, QA-NLI strips away QA-specific formats to focus solely on whether the premise entails the hypothesis, directly measuring intrinsic natural language inference ability.

Stanford/web_questions (Entity Linking)

- **Link:** `https://huggingface.co/datasets/Stanford/web_questions`
- **Explanation:** WebQuestions consists of 6,642 user-posed questions whose answers are Freebase entities, so a system must first detect the mention in each question and disambiguate it to the correct Freebase ID before answering—directly evaluating its entity-linking ability.

Question 2

(a)

1. Self Consistency

Description: Rather than greedily decoding a single chain-of-thought (CoT), we sample N independent reasoning paths and then majority-vote on the final answers to pick the most frequent one — the intuition being that the correct answer tends to recur across diverse valid chains

Advantages: No extra training, annotations, or auxiliary models required—just sample multiple CoT outputs and vote

Bottlenecks:

- Total forward-pass cost scales linearly with N.
- Both the GPU's number-crunching work and its data load/store operations increase in direct proportion to N.

Parallelization: Yes—can batch all N generations in parallel on a single GPU (subject to memory).

2. Choosing Chains Based on Verifiers

Description: Similar to self-consistency, but instead of majority-voting on the final answers, we use a verifier to score each chain and select the one with the highest score. It can be viewed as a search tree where we only deapth-first search nodes that pass the verifier.

Advantages:

- Can be more efficient than self-consistency, as it avoids generating and evaluating all possible chains.
- It also improves correctness by filtering out invalid chains.

Bottlenecks:

- The verifier itself can be a bottleneck, as it needs to be fast and accurate.
- Same N-fold generation cost as self-consistency.

Parallelization: Yes—can batch all N generations in parallel. Since the verification of each chain is independent, we can also parallelize the verification step.

3. Increasing Compute Budget

Description: Rather than running a single large model once, you sample N outputs from a smaller model and then use an (automatic) verifier—unit tests, rule-based checks, or a learned judge—to select the best candidate. Hassid et al. (2024) showed that under equal computational budgets, this can match or exceed a much larger model's performance on code-generation tasks

Advantages:

- We can gain more accuracy with many "weak" models than with a single "strong" model while keeping the same compute budget.
- We can plug in any verifier we want, so we can use a simple rule-based verifier or a more complex learned verifier.

Bottlenecks:

- Success hinges on verifier quality.
- The verifier adds to the overall cost if it is not fast enough.
- We need to wait for all N runs to finish before we can start verifying.

Parallelization: Yes—both sampling and verification can be paralleized. We can run all N generations in parallel on a single GPU (subject to memory). The verification step can also be parallelized, as each verification is independent.

(b)

I'd go with Self-Consistency. It lets you generate N independent reasoning chains in one batch on your GPU and then select the answer that appears most often—no extra models or complex engineering needed. For a truly challenging scientific problem, you're unlikely to have a simple verifier that can check every subtle inference, so building one can be fragile and time-consuming. Majority-voting across multiple chains often helps catch random mistakes, but it's not guaranteed. Still, it gives you a reasonable way to improve confidence without adding significant compute or development overhead.

2 Programming Exercise

Repository URL: `https://github.com/dtahory/anlp_ex1`

Did the configuration that achieved the best validation accuracy also achieve the best test accuracy?

Yes. The configurations achieved test accuracy values very close to their corresponding validation accuracy, maintaining the same ranking order.

- Configuration 1: Validation Accuracy = 0.8529, Test Accuracy = 0.8214
- Configuration 2: Validation Accuracy = 0.8088, Test Accuracy = 0.7826
- Configuration 3: Validation Accuracy = 0.7034, Test Accuracy = 0.6898

Qualitative analysis

I analyzed 5 validation pairs where the best-performing model (cfg_1) correctly predicted the label, but the worst-performing model (cfg_3) did not:

- **Example 1:**
  - **Sentence 1:** Magnarelli said Racicot hated the Iraqi regime and looked forward to using his long years of training in the war.
  - **Sentence 2:** His wife said he was "100 percent behind George Bush" and looked forward to using his years of training in the war.
  - **Label:** 0, **cfg_1:** 0, **cfg_3:** 1
- **Example 2:**
  - **Sentence 1:** The dollar was at 116.92 yen against the yen, flat on the session, and at 1.2891 against the Swiss franc, also flat.
  - **Sentence 2:** The dollar was at 116.78 yen JPY =, virtually flat on the session, and at 1.2871 against the Swiss franc CHF =, down 0.1 percent.
  - **Label:** 0, **cfg_1:** 0, **cfg_3:** 1
- **Example 3:**
  - **Sentence 1:** No dates have been set for the civil or the criminal trial.
  - **Sentence 2:** No dates have been set for the criminal or civil cases, but Shanley has pleaded not guilty.
  - **Label:** 0, **cfg_1:** 0, **cfg_3:** 1
- **Example 4:**
  - **Sentence 1:** "Sanitation is poor ... there could be typhoid and cholera," he said.

  - **Sentence 2:** "Sanitation is poor, drinking water is generally left behind ... there could be typhoid and cholera."
  - **Label:** 0, **cfg_1:** 0, **cfg_3:** 1

- **Example 5:**
  - **Sentence 1:** Friday, Stanford (47-15) blanked the Gamecocks 8-0.
  - **Sentence 2:** Stanford (46-15) has a team full of such players this season.
  - **Label:** 0, **cfg_1:** 0, **cfg_3:** 1

From these examples, I conclude that the lower-performing model (cfg_3) is overly reliant on superficial token overlap and fails to capture small but crucial semantic cues that distinguish these non-paraphrases. Specifically:

Numeric precision: It ignores tiny but meaning-changing number shifts (e.g., 116.92 $\rightarrow$ 116.78, 47-15 $\rightarrow$ 46-15).

Clause insertions: It misses the impact of added qualifiers or after-thoughts (e.g., "but Shanley ... pleaded not guilty," "drinking water ... left behind").

Context shifts: It treats shared long spans (e.g., "looked forward to using his ... training in the war") as a sign of equivalence, even when the subjects or sentiments differ.