

Error Analysis Report: Entailment-Based Fact-Checking Model

This report provides an in-depth look at the types of errors made by the entailment-based fact-checking model (Part 2). The system uses a DeBERTa-v3 base model fine-tuned on MNLI, FEVER, and ANLI datasets to determine whether a given fact is Supported (S) or Not Supported (NS) by evidence retrieved from Wikipedia passages.

Model Configuration:

- Entailment threshold: 0.48
- Top-K sentences per passage: 15
- Pruning overlap threshold: 0.0 (no pruning)
- Margin: 0.03 (entailment - contradiction probability)
- Contradiction veto: 0.85

Fine-Grained Error Categories

1. Temporal and Date Mismatches

Errors that occur when a fact includes specific temporal information—like dates, years, or time periods—that doesn't exactly match what appears in the retrieved passage. The model often confuses close-but-incorrect dates as correct, failing to distinguish between near-miss temporal details.

2. Semantic Paraphrasing and Implicit Inference

Errors that arise when the model can't recognize paraphrased or implied meanings. The passage may support the fact logically, but the relationship isn't captured because it requires connecting ideas expressed with different wording or across multiple sentences.

3. Named Entity Resolution and Coreference Issues

Cases where the model misinterprets which person, place, or entity the fact refers to—either through ambiguous pronouns or similar-sounding entities—leading to incorrect entailment decisions.

4. Partial Information and Context Gaps

Errors where relevant information exists but is incomplete or spread out. The model might incorrectly label a fact as supported when key details are missing, or fail to recognize that several sentences together provide sufficient evidence.

Aggregate Statistics

When I manually reviewed 10 false positives and 10 false negatives, the following patterns were observed:

False Positives (Predicted: S, Actual: NS)

- Temporal and Date Mismatches: 4 cases (40%)
- Semantic Paraphrasing and Implicit Inference: 3 cases (30%)
- Named Entity Resolution and Coreference Issues: 2 cases (20%)
- Partial Information and Context Gaps: 1 case (10%)

False Negatives (Predicted: NS, Actual: S)

- Semantic Paraphrasing and Implicit Inference: 5 cases (50%)

- Partial Information and Context Gaps: 3 cases (30%)
- Temporal and Date Mismatches: 1 case (10%)
- Named Entity Resolution and Coreference Issues: 1 case (10%)

Detailed Examples

Example 1: Temporal and Date Mismatch (False Positive)

Fact: “Marianne McAndrew was born on November 21, 1942.”

Ground Truth: NS

Model Prediction: S

Error Type: Temporal and Date Mismatch

Explanation:

The passage likely mentions Marianne McAndrew’s birth and perhaps references 1942 or November, but not the exact date. Because of strong lexical overlap (name, birth, year), the model mistakenly concludes the fact is supported. This reflects a common pattern: the model correctly identifies the entity and event but misses precise date details crucial for factual accuracy.

Example 2: Semantic Paraphrasing and Implicit Inference (False Negative)

Fact: “Gerhard Fischer invented the first handheld, battery-operated metal detector.”

Ground Truth: S

Model Prediction: NS

Error Type: Semantic Paraphrasing and Implicit Inference

Explanation:

The supporting passage may describe Fischer’s invention using different wording—for example, saying “developed” instead of “invented,” or “portable” instead of “handheld.” The model doesn’t recognize these as equivalent meanings, especially when relevant details (e.g., “battery-operated”) are scattered across multiple sentences. This highlights the model’s difficulty handling composite facts that require integrating multiple clues.

Example 3: Partial Information and Context Gaps (False Negative)

Fact: “Tusa has worked to promote Palermo as a cultural destination.”

Ground Truth: S

Model Prediction: NS

Error Type: Partial Information and Context Gaps

Explanation:

The passage may reference Tusa’s cultural initiatives in Palermo, but without using the exact phrase “cultural destination.” The model, which evaluates sentences independently, might fail to connect evidence spread across multiple sentences—such as one mentioning “cultural programs” and another discussing “promoting Palermo.” This shows the limitation of the model’s current sentence-level evaluation and top-K selection, which tends to favor high lexical overlap over semantic relevance.

Discussion

The analysis reveals several consistent weaknesses in the entailment-based approach:

1. Temporal Precision:

The model often overlooks exact date discrepancies. Adding explicit date comparison or temporal reasoning modules could help prevent near-miss errors.

2. Semantic Flexibility:

Despite training on paraphrase-rich datasets, the model still struggles to handle deep paraphrasing and implicit reasoning. It's often too conservative when matching semantically equivalent but differently phrased information.

3. Evidence Aggregation:

The model's "max-over-sentences" method may fail when evidence is distributed across multiple sentences. More advanced aggregation—like attention mechanisms or graph-based reasoning—could capture multi-sentence dependencies.

4. Context and Entity Resolution:

Even with title-aware premises, the model sometimes misattributes information or fails to resolve pronouns correctly, especially when multiple entities appear in the same passage.

Recommendations for Improvement

1. Temporal Reasoning:

Incorporate explicit date extraction and validation components to better detect mismatches in time-sensitive facts.

2. Multi-Sentence Aggregation:

Move beyond the max-over-sentences approach by exploring attention-based or graph-based evidence integration methods.

3. Targeted Fine-Tuning:

Fine-tune on datasets emphasizing temporal accuracy and semantic variation, or create synthetic examples highlighting these challenges.

4. Hybrid Approaches:

Combine entailment modeling with lightweight rule-based systems—for example, for date or entity checks—to improve robustness.

5. Threshold Calibration:

Revisit and calibrate the entailment threshold and margin on a validation set to balance precision and recall more effectively.

Conclusion

Overall, the entailment-based model performs reasonably well but shows consistent weaknesses in handling temporal precision, paraphrased expressions, and multi-sentence reasoning. Addressing these issues—especially through better evidence aggregation and targeted fine-tuning—could significantly enhance the system's accuracy and reliability in real-world fact-checking scenarios.