Context-Aware Plagiarism Detection

Evelyn Yee - CS 329T

github

Background

Plagiarism is a classic issue in education, and it has **only gotten easier** with the development of LLMs.

- Difficult **false positives**: critique, citing, reference
- Difficult **false negatives**: paraphrasing, summary

Our goal: **Empower** instructors to more effectively identify plagiarism in student works.

To do this, a plagiarism detection system needs to:

- **run quickly** on documents of different sizes and in different contexts
- **dynamically adapt** over time, as students make new submissions and the corpus of relevant source documents also changes
- **provide verifiable evidence,** allowing instructors to make informed assessments of student work
- **consider nuance** in the provided literary context

Dimensions of Trust

Grounding

- Plagiarism accusations need to be grounded in evidence from **both the student's submission and the source documents.**
- Grounding gives instructors **specific evidence** to evaluate, expediting the evaluation process.

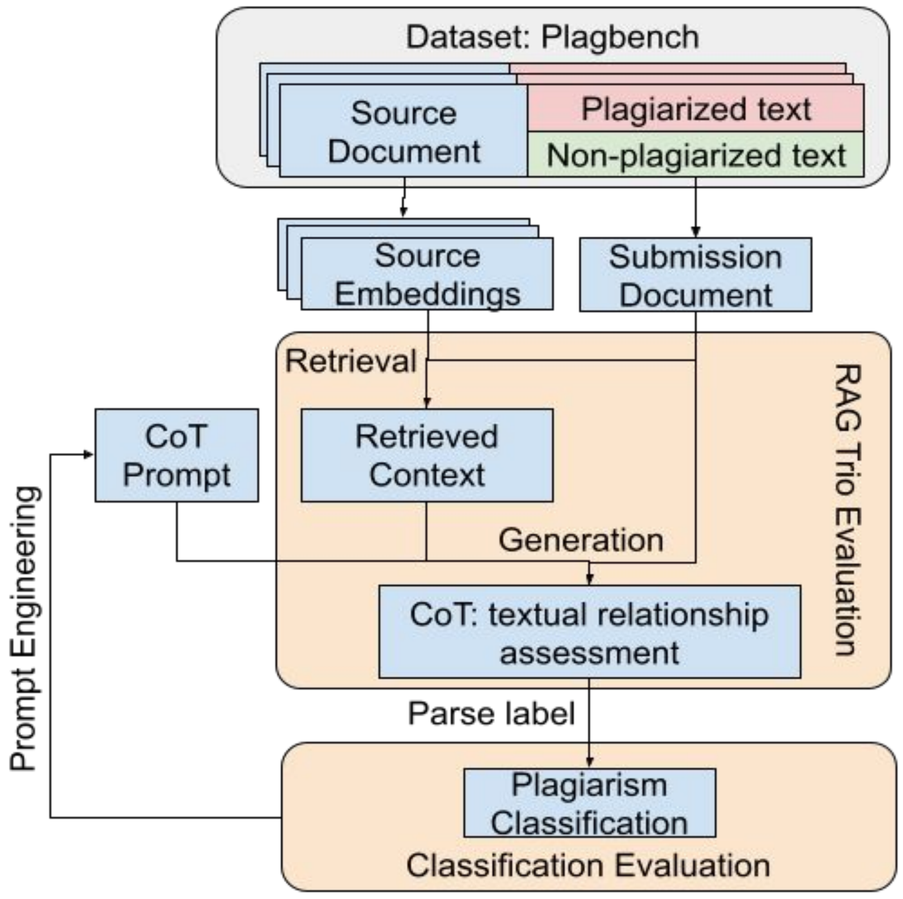
Interpretability

- A **human-readable** Chain of Thought report allows the instructor to understand the assessment, rather than having to blindly accept the score..

Confidence

- The system acknowledges when it is less confident, allowing instructors to make i**nformed choices** about how to handle the result.

Pipeline



Jooyoung Lee, Toshini Agrawal, Adaku Uchendu, Thai Le, Jinghui Chen, and Dongwon Lee. 2024. Plagbench: Exploring the duality of large language models in plagiarism generation and detection

Baselines

- **Text Heuristics**: logistic regression on n-gram maximum similarity metrics
- **Parametric Knowledge**: Chain of Thought without access to source documents
- **Direct Prompting**: RAG with non-Chain of Thought prompt

Performance

Classification Performance

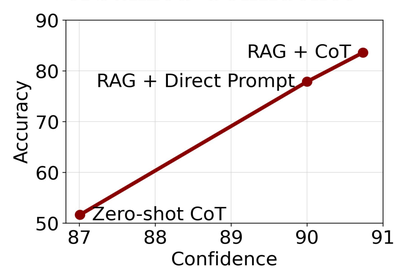
Pipeline	Precision	Recall	Accuracy	
			Fine	Coarse
N-gram Heuristics	100.00	83.05	87.70	91.80
Zero-shot CoT	50.00	1.69	0.00	51.64
RAG + Direct Prompt	68.60	100.00	35.59	77.87
RAG + CoT	74.68	100.00	52.50	83.61

RAG Trio

Pipeline	Context Relevance	Groundedness	Answer Relevance
Zero-shot CoT	–	46.76	89.34
RAG + Direct Prompt	86.07	79.91	69.67
RAG + CoT	85.25	67.79	100.00

Evaluated using the TruLens framework, with custom criteria for the plagiarism detection task.

Verbalized Confidence



Conclusions

- RAG offers a **lightweight**, **scalable**, and **customizable** approach to plagiarism detection
  - Retrieval and Chain of Thought are essential to the system performance
- For this task, **n-grams are accurate** if you have definite access to the single source document
  - lower interpretability + confidence
- **Interpretability is key** for real-life deployment
  - promote seamless integration between humans and trustworthy AI systems