

Plagiarism Detection Using Human-in-the-Loop AI

Manqian Liao, Sinon Tan, Mirza Basim Baig

Duolingo

April 15, 2023

Annual Meeting of National Council on Measurement in Education



Outline

- Background & Motivation
- The Proposed Plagiarism Detection System
- Development Process & Design Considerations
- Quality Assurance
- Discussion

Background & Motivation

What is plagiarism?

- Plagiarism is widely perceived as a form of academic misconduct or moral transgression
- In the field of second-language (L2) writing, there is controversy around the definition of plagiarism (Pecorari & Petrić, 2014)

Plagiarism



Intertextual
Relationship

What is plagiarism?

- Definition of plagiarism in this presentation:

Plagiarism



**Intertextual
Relationship**

- intentional cheating behavior
- violate testing rules and use external resources on one's own response

- a linguistic phenomenon
- overlap between a new text and a source text

Motivation

Literature Gap / Challenge

-
- 1 Existing plagiarism detection methods in language assessment (e.g., Wang, Evanini, Mulholland, et al., 2019) do not explicitly distinguish between **deceptive** and **non-deceptive** intertextual relationships
-
- 2 There is a disconnect between **research-stage tools** (Donaldson et al., 1981; Si et al., 1997; Wang, Evanini, Mulholland, et al., 2019; Wang, Evanini, Qian, et al., 2019) and a **production-ready plagiarism detection system**
-

Motivation

We intend to develop a production-ready plagiarism detection system that meets the following goals:

- Distinguish deceptive and non-deceptive intertextual relationships in order to address the issue of unfair penalization for test-takers from certain language backgrounds
- Bridge the disconnect between existing research-stage tools and a usable production-ready plagiarism detection system

The Proposed Plagiarism Detection System

Rationale

Literature Gap / Challenge

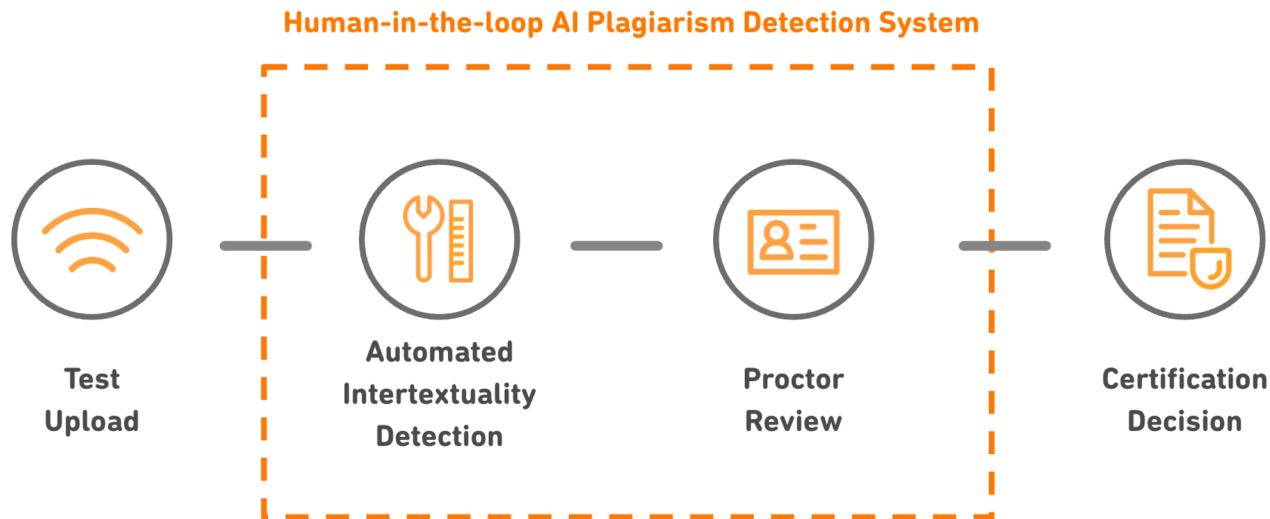
Solution

Lack of distinction between deceptive and non-deceptive intertextual relationships	→ Utilizing human-in-the-loop artificial intelligence (AI)
Disconnect between research-stage tools and a production-ready plagiarism detection system	→ Optimizing system usability following best practices recommended by review on plagiarism detection systems, explainable AI and responsible AI guidelines (Arrieta et al., 2019; Burstein, 2023; Dignum, 2017; Foltýnek et al., 2020; Shneiderman, 2021),

Properties of the Proposed System

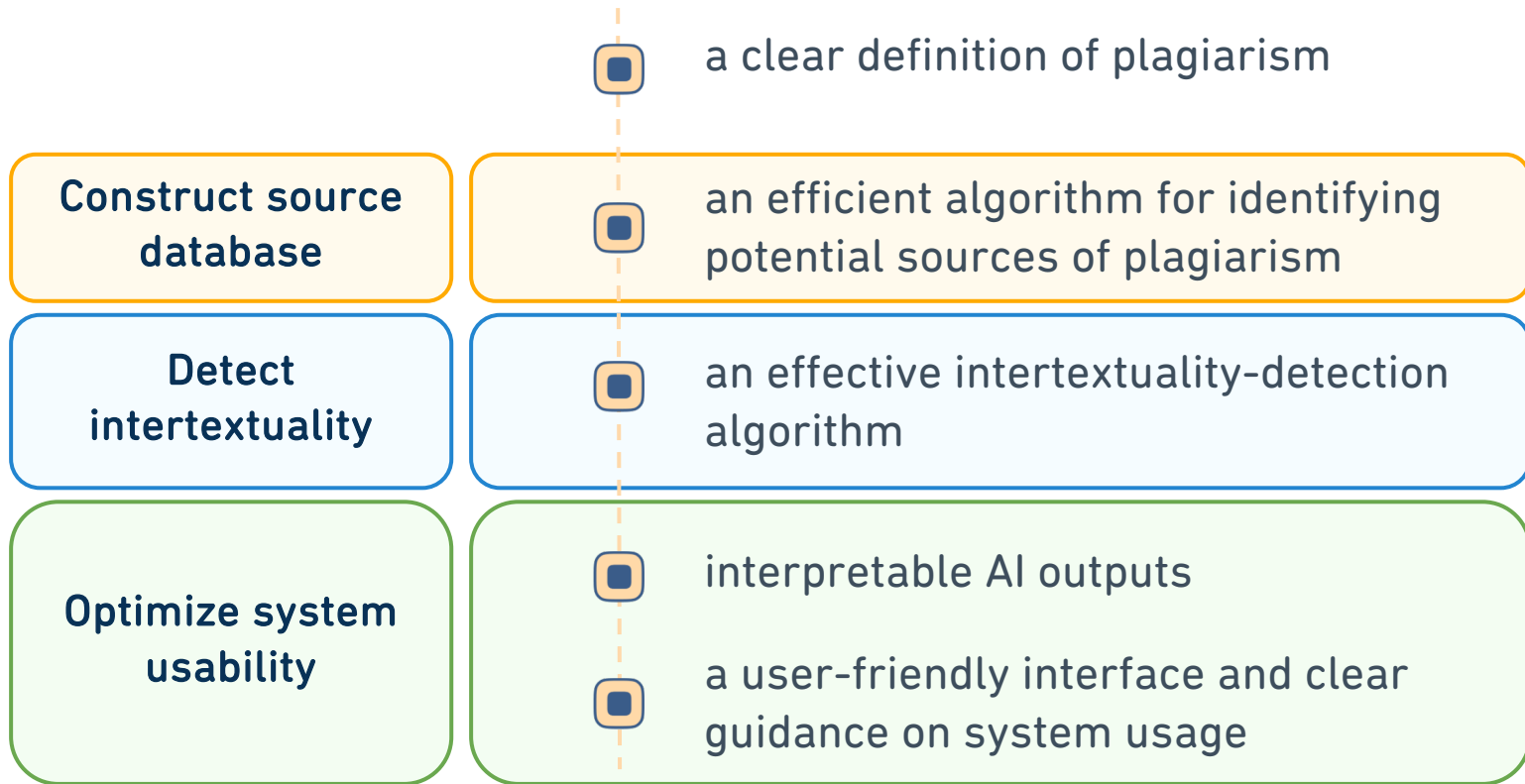
-  a clear definition of plagiarism
-  an efficient algorithm for identifying potential sources of plagiarism
-  an effective intertextuality-detection algorithm
-  interpretable AI outputs
-  a user-friendly interface and clear guidance on system usage

Overview of Test Certification Process



Development Process & Design Considerations

Building Blocks of the System



Construct source database

Internet content database

- Plagiarism sources that may have been accessed via Internet
- Useful if a test taker used a search engine during the test



historical response database

- Historical test takers' responses
- Useful if a test taker used a private braindump shared by an earlier test taker



Detect intertextuality

- Intertextuality Detection Methods
 - N-gram fingerprinting, bag-of-word, edit distance
 - Goal: classify whether there was a significant intertextual relationship between the response and the source
- Benchmark dataset
 - N=906 (868 plagiarized, 38 not plagiarized)
- Evaluation measures
 - Recall: the proportion of plagiarized responses that were classified as "positive"
 - Precision: the proportion classified as "positive" that were actually plagiarized
 - Cohen's Kappa

Detect intertextuality

● Classification Results

Intertextuality	# of true	# of false			
detection method	positives	positives	recall	precision	Cohen's Kappa
n-gram fingerprinting	828	1	0.95	1.00	0.62
bag-of-word	795	2	0.92	1.00	0.46
edit distance	802	16	0.92	0.98	0.31

**Optimize
system usability**

Present technical results

Enhance Visuals

Provide system usage guidelines

Decision 1: which statistics should be displayed to convey the extent of overlap between the response and the source text?

- Define a text overlap statistic that is easy for the proctors to understand and has good face validity
- In n-gram fingerprinting method, use **the percentage of matched n-grams between response and source** as the text overlap statistic

**Optimize
system usability**

Present technical results

Enhance Visuals

Provide system usage guidelines

Decision 2: which responses should be escalated to the proctors for further review?

- Define an escalation threshold based on the text overlap statistic
- Ideally, set the escalation threshold to maximize the chance of discovering plagiarism while avoiding overburdening the proctor

Development Process & Design Considerations

Optimize
system usability

Present technical results

Enhance Visuals

Provide system usage guidelines

Similar sentences
are highlighted

Text Overlap Statistics

Similar historical
test sessions

Similar Internet
content

Not Plagiarism

Plagiarism

Shortcuts for proctor decisions

The screenshot displays a web interface titled "Similar Texts Found!". At the top, it shows "Item Type:" and "Max similarity: 60.00%". Below this, there are two main columns: "Current Response" and "Past exam:". The "Current Response" column contains a paragraph of text with several sentences highlighted in yellow. The "Past exam:" column contains a similar paragraph, also with some text highlighted. To the right of these columns, there is a "Similar Results:" section listing three items with their respective similarity percentages: 60.00%, 62.27%, and 49.09%. At the bottom of the interface, there are two buttons: "Not Plagiarism" and "Plagiarism".

Development Process & Design Considerations

Optimize
system usability

Present technical results

Enhance Visuals

Provide system usage guidelines

- System usage guidelines include:
 - Rules of thumb for plagiarism decision
 - Emphasis of the distinction between an intertextual relationship and plagiarism
 - Examples

Similar Texts Found! Max Similarity: 66.67%

Item Type: Max similarity: 66.67% Current similarity: 66.67%

Prompt:

Current Response

One of the best feelings in the world that also brings so much of cheer to life is laughter. It really is one of the best medicines in the world. Also, whether it is a smile or just a slight giggle, laughter completely alters the atmosphere and mood of the surroundings. Furthermore, it makes you feel good and everyone around you will also feel positive vibes. The laughter is the best medicine. Yes, I will definitely agree with the statement. I would like to share an incident, recently one of my friend was depressed and not coming to work. So we all friends planned to get him off from his depression. Finally, we made him laugh and he is out of his depression and happy. This was the situation I felt laughter is the best medicine in the world. There are also many health benefits by laughing.

Past exam:

One of the best feelings in the world that also brings so much of cheer to life is laughter. It really is one of the best medicines in the world. Also, whether it is a smile or just a slight giggle, laughter completely alters the atmosphere and mood of the surroundings. Furthermore, it makes you feel good and everyone around you will also feel positive vibes. Laughter acts as a powerful antidote for pain, stress, and conflict.

An example response that is determined as plagiarism

Similar Texts Found! Max Similarity: 38.46%

Item Type: Max similarity: 38.46% Current similarity: 38.46%

Prompt:

Current Response

The advantage of to learn through personal experience is that¹ we can keep them in mind. The disadvantage is that² it is not general. The advantage of to learn through the³ advice is that it is general. The disadvantage is that⁴ if we have no friend, we can not get any advices.

Past exam:

The advantage of learn through personal experience is that¹ people could have their own perspective and become unique. And the disadvantage is that² people would have less ability to encourage others. The advantage of learn through the³ advice of others is people could become trendy and well hang out with others. The disadvantage is that⁴ people would imitate almost everything of other people.

An example response that is determined as not plagiarism

Quality Assurance

Quality Assurance

Goal

Measure

Keep source database up to date



Update the source database periodically

Monitor the impact on stakeholders



Track the percentage of flagged test takers

Evaluate the performance of AI



Track the AI classification precision

Evaluate the human component



Assess the consistency of certification decisions among proctors using proctor calibration tool (Belzak et al., 2023)

Ensure the data are used appropriately



Store and use data following data protection policies that are compliant to General Data Protection Regulation (GDPR)

Discussion

Summary of System Design Considerations

In designing the system, we strived to maintain three balances:

- balance the need for test security with the need for a positive test-taker experience
- balance the performance and transparency of the AI system
- balance the power of discovering plagiarism and the need to prevent proctor burnout

Future Directions

- Incorporate large language models (LLM) such as GPT-4 (OpenAI, 2023), e.g.,
 - Reduce the number of false positives outputted from the AI by using LLM to identify intertextual relationships that are likely to be plagiarism
 - Use LLM to assist proctors in making more efficient decisions when reviewing potential cases of plagiarism

References

- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI (arXiv:1910.10045). arXiv. <http://arxiv.org/abs/1910.10045>
- Burstein, J. (2023). Responsible AI Standards. <https://duolingo-papers.s3.amazonaws.com/other/DET+Responsible+AI+033123.pdf>
- Dignum, V. (2017). Responsible artificial intelligence: Designing AI for human values.
- Donaldson, J. L., Lancaster, A.-M., & Sposato, P. H. (1981). A plagiarism detection system. Proceedings of the Twelfth SIGCSE Technical Symposium on Computer Science Education, 21–25.
- Foltýnek, T., Dlabolová, D., Anohina-Naumeca, A., Razi, S., Kravjar, J., Kamzola, L., Guerrero-Dib, J., Çelik, Ö., & Weber-Wulff, D. (2020). Testing of support tools for plagiarism detection. International Journal of Educational Technology in Higher Education, 17(1), 46. <https://doi.org/10.1186/s41239-020-00192-4>
- OpenAI. (2023). GPT-4 Technical Report. <https://cdn.openai.com/papers/gpt-4.pdf>
- Pecorari, D., & Petrić, B. (2014). Plagiarism in second-language writing. Language Teaching, 47(3), 269–302. <https://doi.org/10.1017/S0261444814000056>
- Shneiderman, B. (2021). Responsible AI: Bridging from ethics to practice. Communications of the ACM, 64(8), 32–35.
- Si, A., Leong, H. V., & Lau, R. W. (1997). Check: A document plagiarism detection system. Proceedings of the 1997 ACM Symposium on Applied Computing, 70–77.
- Wang, X., Evanini, K., Mulholland, M., Qian, Y., & Bruno, J. V. (2019). Application of an Automatic Plagiarism Detection System in a Large-scale Assessment of English Speaking Proficiency. Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, 435–443. <https://doi.org/10.18653/v1/W19-4445>
- Wang, X., Evanini, K., Qian, Y., & Zechner, K. (2019). Using Very Deep Convolutional Neural Networks to Automatically Detect Plagiarized Spoken Responses. 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 764–771.

Thanks! Questions?

mancy@duolingo.com

