ECIR 2023

Springer

Towards Effective Paraphrasing for Information Disguise

Anmol Agarwal [Ŧ*], Shrey Gupta [Ŧ*], Vamshi Bonagiri [Ŧ], Manas Gaur [§], Joseph Reagle [□] & Ponnurangam Kumaraguru [Ŧ]

**Authors contributed equally*

[Ŧ] IIIT Hyderabad, India
[§] The University of Maryland, Baltimore County
[□] Northeastern University

Background

- Researchers dealing with public user-generated content often need to paraphrase content related to sensitive topics before making it public.

- Posting on Reddit is NOT automatic consent for public distribution
  - Especially applicable for content meant for specific forums
    - Example: Depression, Suicide, Mental Health, Abuse
  - Unwanted attention and publicity

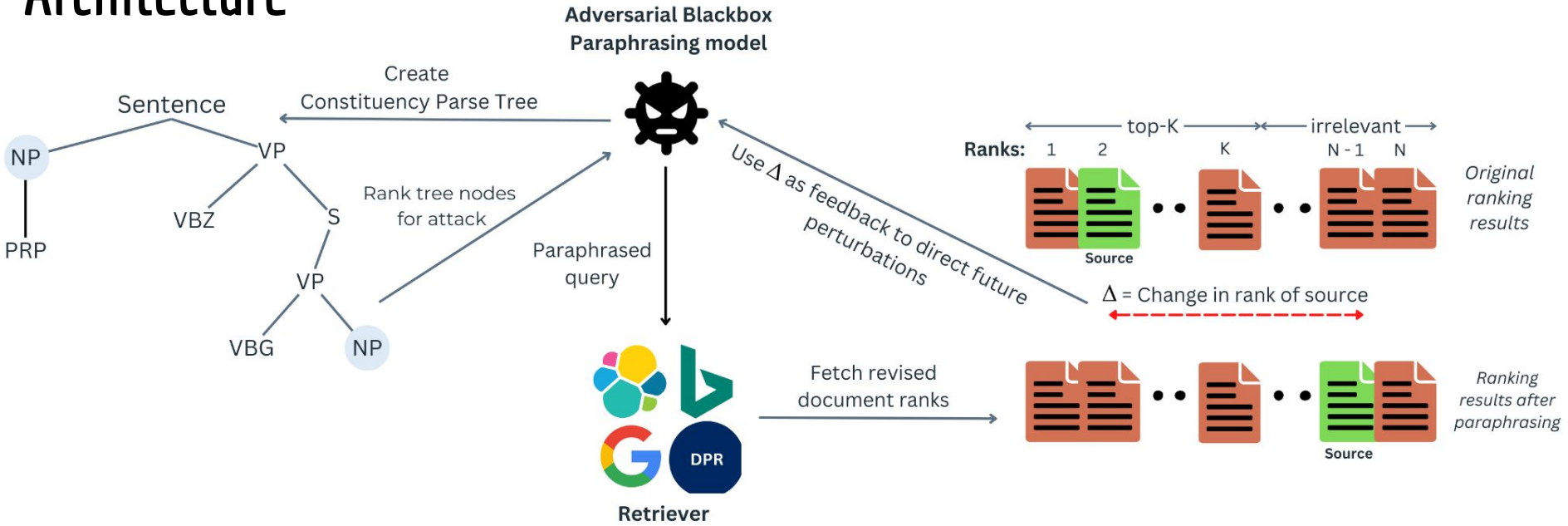
- Shadow Profiling
  - Attention not limited to one community or platform
  - Can spread to others using your username
  - Usernames reflect characteristics and habits

Motivation and our Problem Statement

- Existing AI word spinners (eg. SpinRewriter, WordAI etc.)
  - Ineffective for paraphrasing
  - Sources of paraphrased content are still locatable on search engines[1].
- **Introducing:** an **unsupervised *black-box*** adversarial framework to paraphrase content such that querying snippets of text from it on search engines does not lead back to the original content on the web.
- Given a sentence '**s**,' derived from a document "D", we paraphrase the sentence with 2 aims:
  - **Non-locatability**: Sentence's source "D" is non-locatable
  - **Fidelity**: Semantic meaning of the sentence "s" is preserved
- Our setup:
  - *Retriever used*: Dense Passage Retriever (DPR)
  - *Retriever document store:* 2000 posts from the subreddits r/AmItheAsshole and r/AmItheButtface
  - *Queries*: Single sentences within the posts which result in their source post being within top-2 documents when queried.

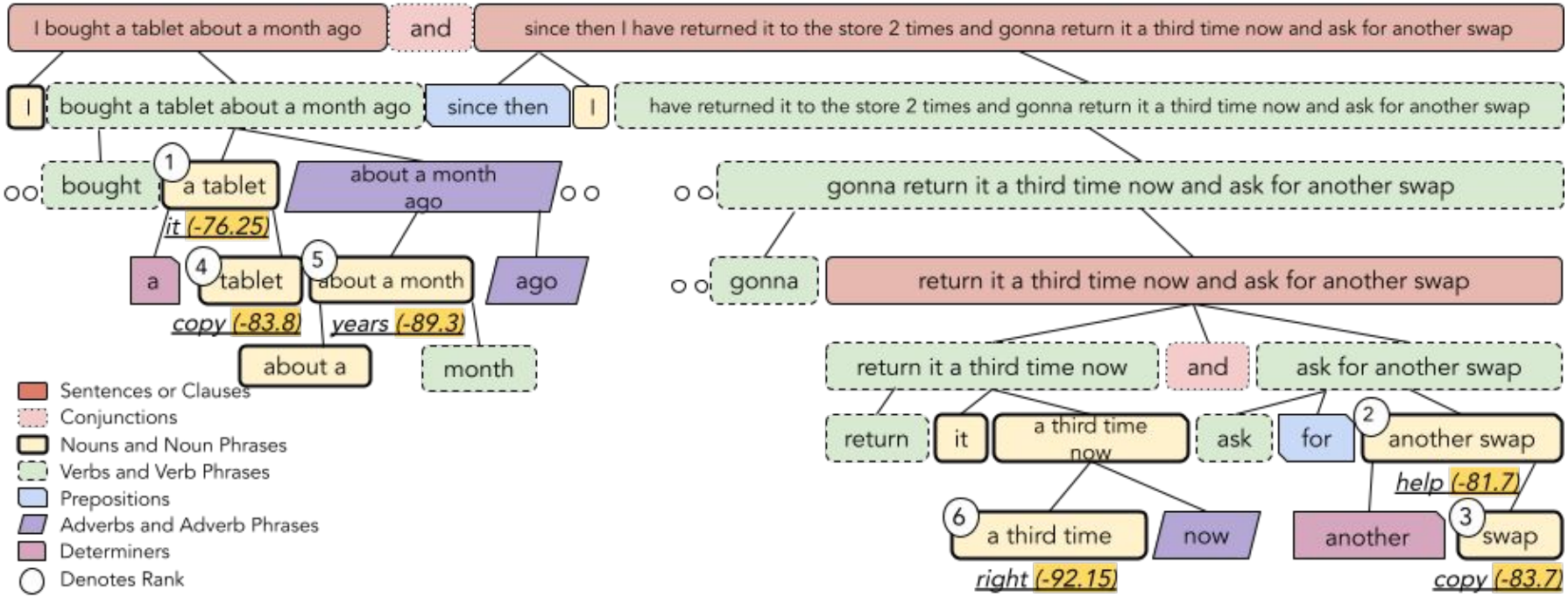
[1]: Reagle, J. and Gaur, M. 2022. Spinning words as disguise: Shady services for ethical research?. First Monday, vol. 27, no. 1, Jan. 2022

Architecture



STEP 1: Shortlisting parts of the sentence eligible for paraphrasing
STEP 2: Ranking the shortlisted parts of the sentence to prioritize our attack
STEP 3: Creating possible paraphrases of the part of the sentence to attack
STEP 4: Repeating steps 1-3 to expand our approach for multi-phrase paraphrasing guided by feedback from previous paraphrasing attempts

STEP 1: Shortlisting parts of the sentence eligible for attack



Sentence "s": I bought a tablet about a month ago and since then I have returned it to the store 2 times and gonna return it a third time now and ask for another swap.

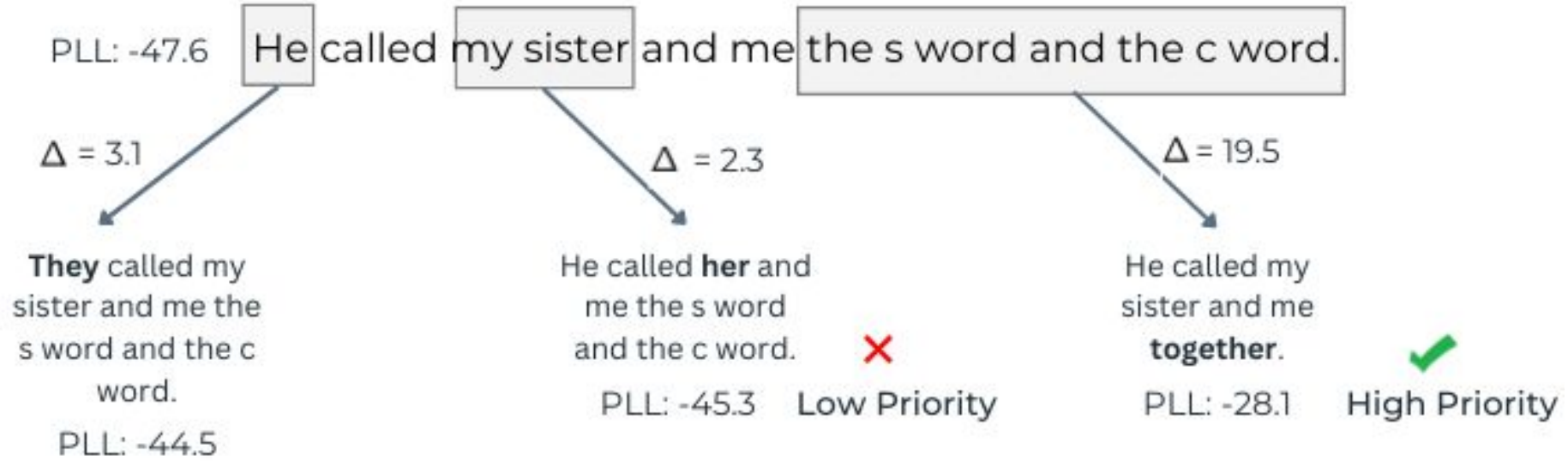
- Substrings such as "and since", "times and gonna" alone DO NOT have independent meaning.
- We shortlist only those substrings which are present in some node of the Constituency Parse Tree *.

** Using the Berkeley Neural Parser*

STEP 2: Ranking shortlisted nodes to prioritize our attack

We prioritise which parse tree nodes to attack using *Pseudo log-likelihood (PLL)* scores.

PLL: Probability of a sentence from BERT, by iteratively masking every word in the sentence and then summing the log probabilities.



Replacing the phrase “the s word and the c word” with “together” leads to maximum increase in the sentence occurrence probability. This indicates that “the s word and the c word” is the *most peculiar part* of the sentence and hence, a potent location for attack.

NOTE: *The replacement phrase here is JUST for ranking nodes and NOT the actual replacement for the phrase.*

STEP 3: Generating potential paraphrases for top-ranked nodes

Attacking by generating replacements using a combination of:

1) **BERT masked language model :** maintains grammar; takes surrounding content into account; independent of the phrase being replaced

2) **Synonyms in Counter-fitting vector space:** depends on phrase being replaced; does not take surrounding context into account; decreases grammar quality

---

He called me and my sister the s word and the c word.

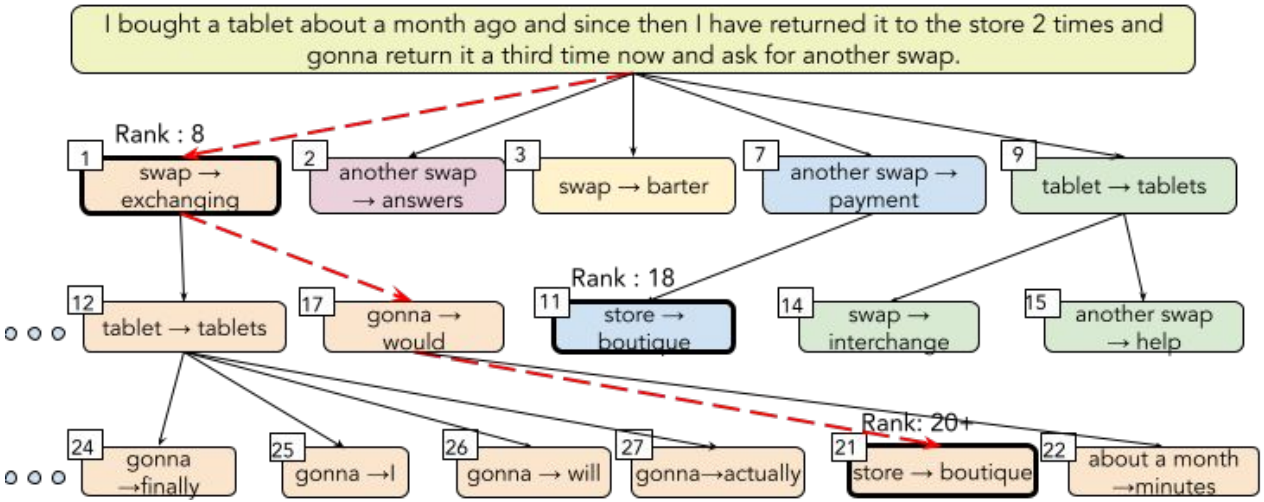
Bert suggestions: friend, father, mother, brother

Counter-fitting vector based suggestions: sibling, hermano, nun, sis

BERT suggestions:

- He called me and my sister names.
- He called me and my sister crazy.
- He called me and my sister out.

STEP 4: Expanding to Multi-Phrase paraphrasing using Beam Search



$$f(s_{paraphrased}) = (1-\alpha) * \underbrace{Sim(s_{org}, s_{paraphrased})}_{\text{semantic similarity (distance from origin)}} + \alpha * \underbrace{\frac{(Rank(s_{paraphrased}, D_{source}) - 1)}{20}}_{\text{non-locatablity of source (estimated distance to target)}}$$

Performance

We succeed in **disguising 82% of the queries** (source document outside top-20) when there are 3 beam levels and 5 nodes per parse tree are expanded.

HR@K	Level 1	Level 2	Level 3
K=1	0.18	0.06	**0.04**
K=5	0.46	0.17	**0.10**
K=10	0.60	0.24	**0.13**
K=20	0.71	0.34	**0.18**

A controlled attack (ie attacking limited number of parse tree nodes) during multi-phrase perturbation
is **MUCH MORE effective** than
a brute force attack during single-phrase perturbation.

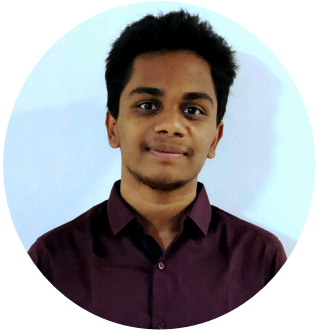
Potential areas for future work

- Including a grammar quality score in the ranking metric to prevent grammatical errors.
- Reducing the number of requests made to the retriever before a successful paraphrasing attempt

The Authors

Anmol Agarwal

Shrey Gupta



Vamshi Krishna

Manas Gaur

Ponnurangam Kumaraguru (PK)

Joseph Reagle

Thank You !

Questions ?

Code Repository:
https://github.com/idecir/idecir-Towards-Effective-Paraphrasing-for-Information-Disguise