Towards Effective Paraphrasing for Information Disguise

Anmol Agarwal*, Shrey Gupta*, Vamshi Bonagiri, Manas Gaur, Joseph Reagle, and Ponnurangam Kumaraguru

IIIT Hyderabad, KAI2 Lab @ UMBC, Northeastern University

ECIR 2023

C2S2

Research center

Motivation and Problem Statement

Researchers dealing with public user-generated content often need to paraphrase content related to sensitive topics like health, violence, drug use, etc, before making it public.

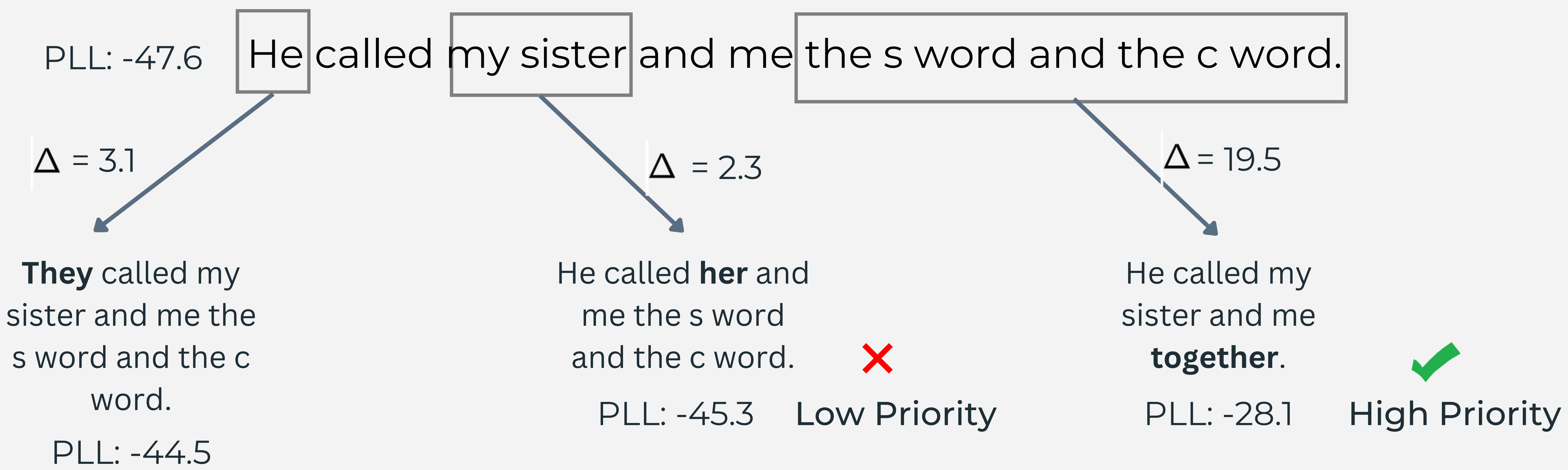
- Existing AI-based automated word spinners (e.g., SpinRewriter, WordAI) are often ineffective as their paraphrased content is still locatable on search engines.
- **Introducing:** an *unsupervised black-box adversarial framework* to paraphrase content such that querying snippets of text from it on search engines does not lead back to the original content on the web.

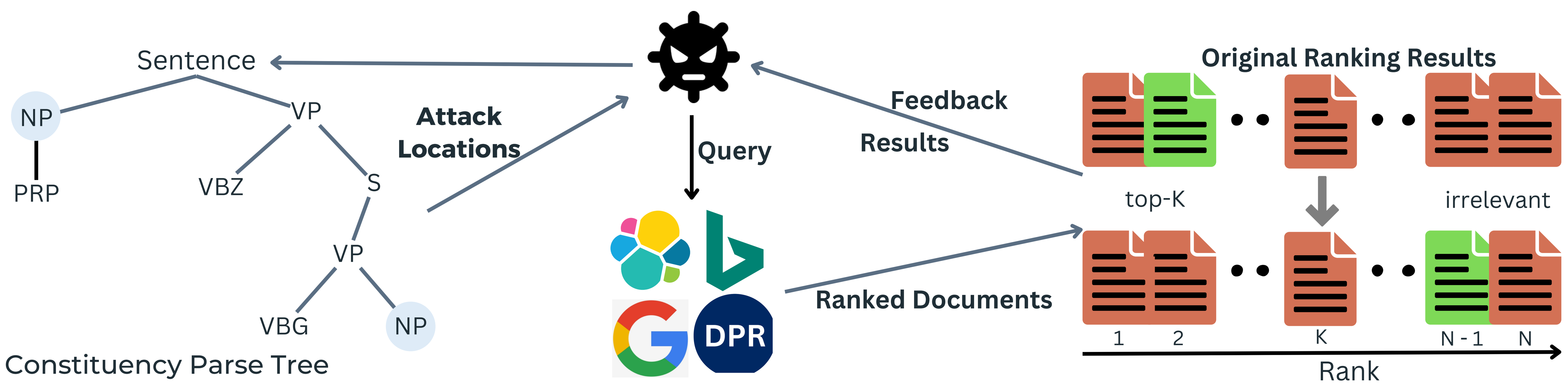
Given a sentence 's,' we paraphrase the sentence with the aim:

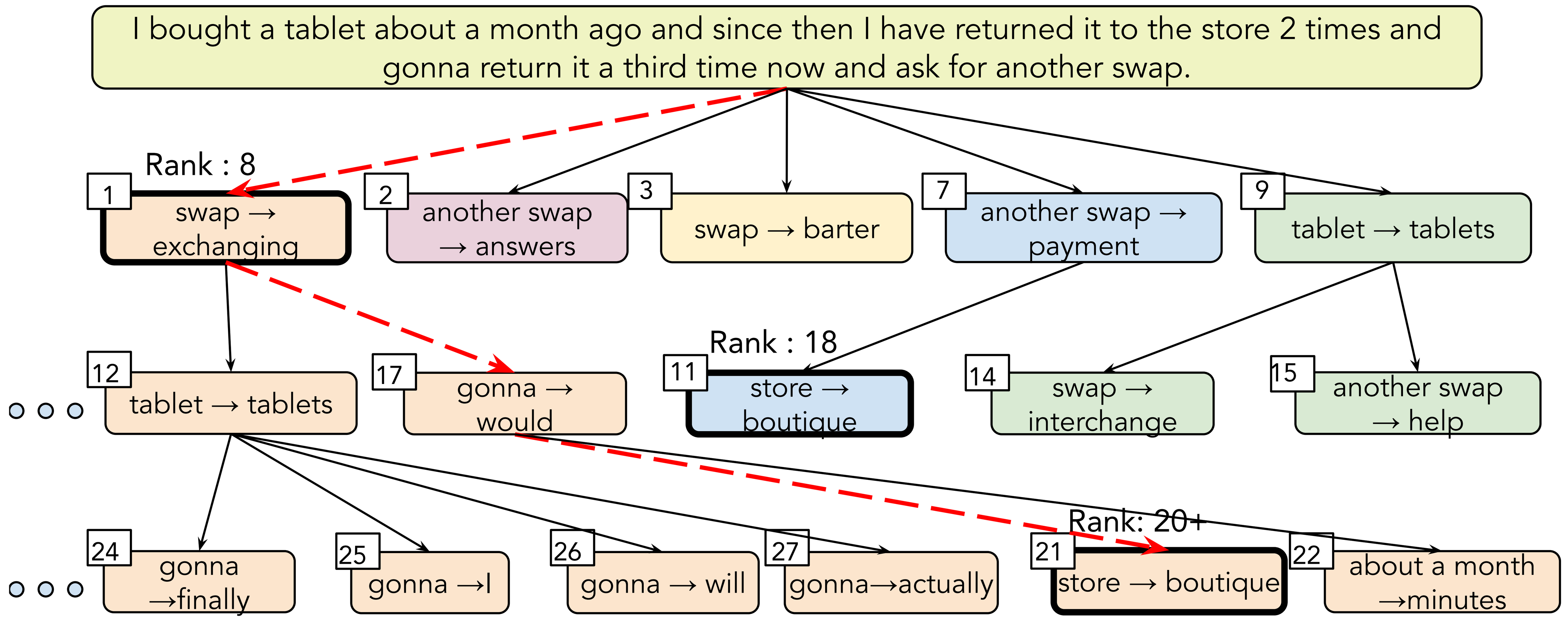
- **Non-locatability**: the sentence's source is not in the top-K results when the sentence is queried on search engines
- **Fidelity**: the semantic meaning of the sentence is preserved

Which part of the sentence to attack first ?

STEP 1: Create a constituency-based parse tree for the sentence.
STEP 2: Prioritise parse tree nodes to attack based on PLL scores.
STEP 3: Rank candidates based on *PLL score difference* after replacing masks with BERT suggestions.
PLL: Probability of a sentence from BERT, by iteratively masking every word in the sentence and then summing the log probabilities.







Types of Attack

Attacking by generating replacements using a combination of:
1) **BERT masked language model** : maintains grammar; independent of the phrase being replaced
2) **Synonyms in Counter-fitting vector space**: depends on phrase being replaced; decreases grammar quaility

Multi-level attack for multi-level perturbations

Expanding single-phrase attack to multiple levels using Beam Search.

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

Results

We succeed in disguising 82% of the queries when there are 3 beam levels and 5 nodes per parse tree are expanded.

References

- Reagle, J. and Gaur, M. 2022. Spinning words as disguise: Shady services for ethical research?. First Monday, vol. 27, no. 1, Jan. 2022
- Jin Yong Yoo and Yanjun Qi. 2021. Towards Improving Adversarial Training of NLP Models. EMNLP 2021