Exploiting Natural Language Services: A Polarity Based Black-box Attack

Fatma GUMUS, M. Fatih AMASYALI

Frontiers of Computer Science, DOI: 10.1007/s11704-021-0198-y

Problems & Ideas

- Problem of exploting NLP services in black-box setting:
  - Querying the victim model to find the important words to manipulate potentially cause a big request-response overhead.
- Ideas: Replace words by their class polarity scores calculated on an external dataset.

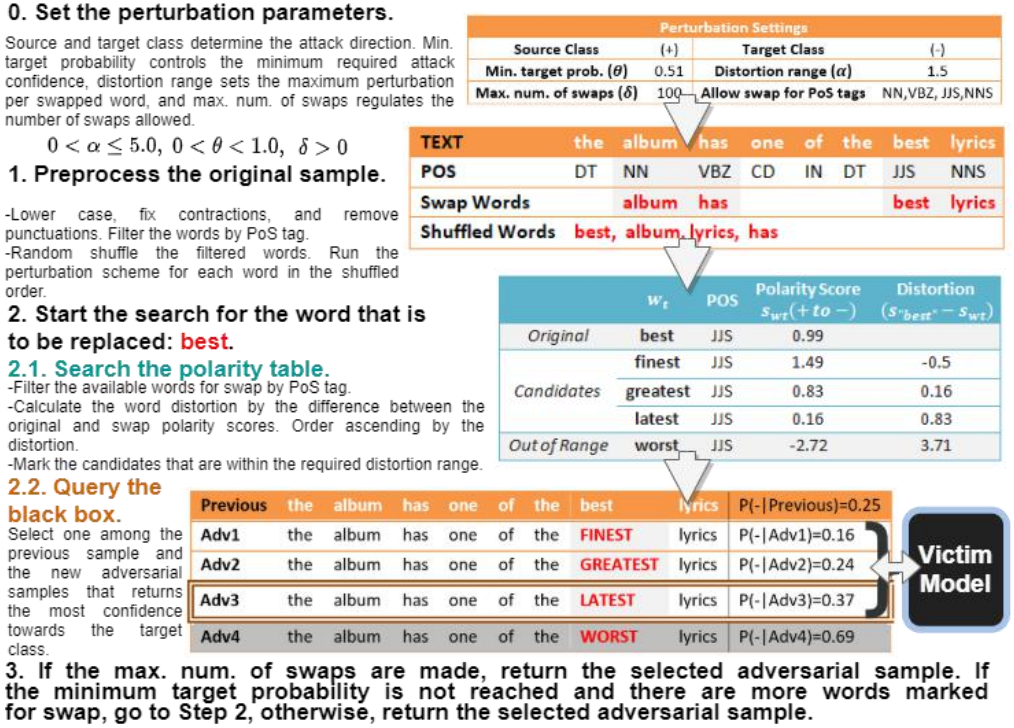

Figure 1. Example on one iteration of our proposed attack.

Main Contributions

- We perform black-box word perturbations without an oracle, or a trained generator.
- Classifiers were fooled by a small number of word replacements.
- The generated samples retain semantic similarity with the originals based on embedding space.
- We evaluate a true black-box attack on IBM Watson Natural Language Understanding (NLU) service to further confirm its applicability on real case scenarios.

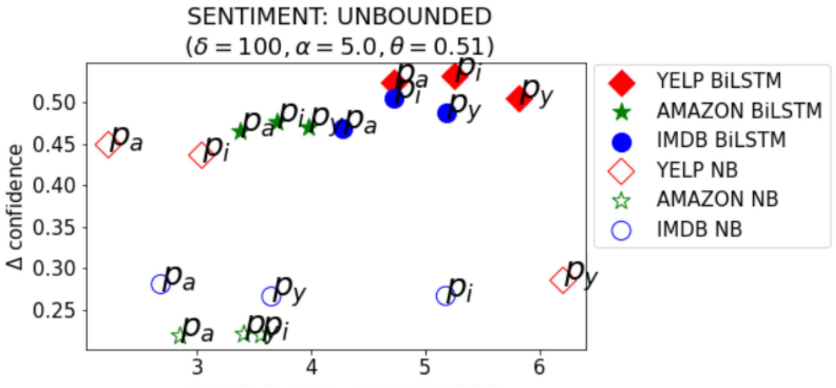

Figure 2. Number of Word replacements (x axis) and the change of probability score on false class (y-axis) on the Sentiment Task.

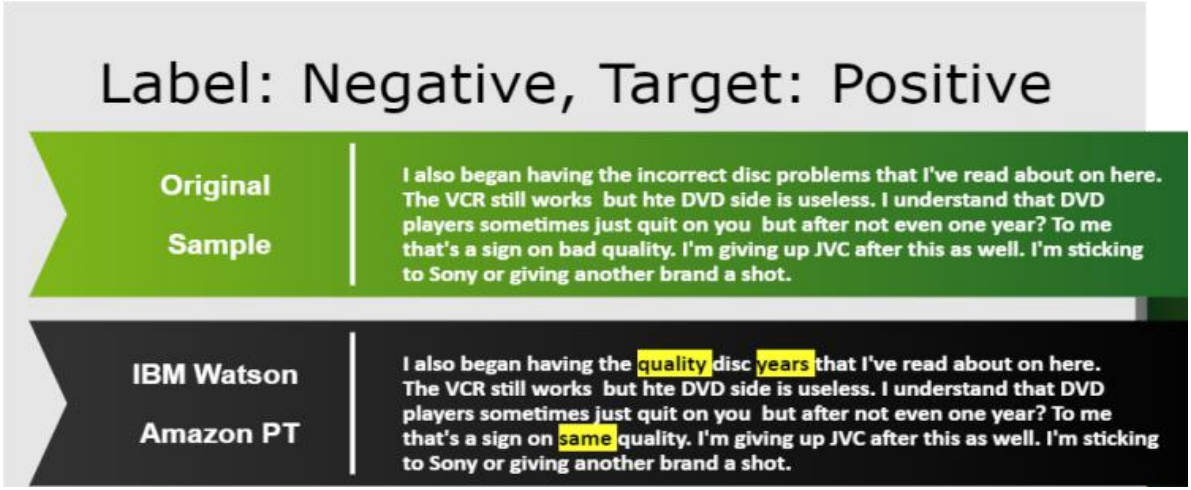

Figure 3. The original and the corrupted sample using Amazon polarity table on IBM Watson NLU victim.