# AI技術を悪用したサイバー攻撃の脅威とその対策

## ～攻撃者の視点から考える～

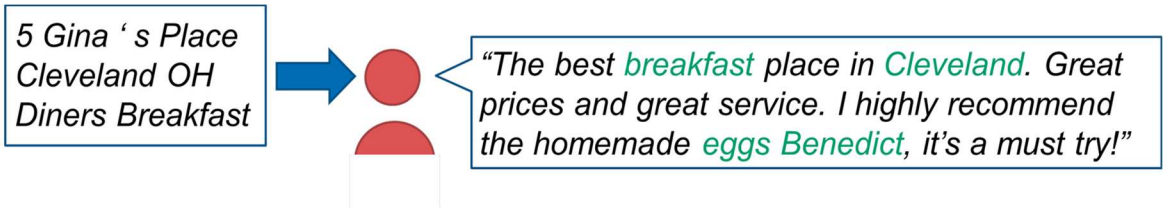
情報科学研究科

○准教授　森川智博（もりかわともひろ）

**キーワード**　**レストラン、レビュー・コメント、評判情報、人工知能技術、攻撃者**

**研究概要**

レストラン等を検索するための評価ランキングサイトにおいて、自らのレストラン評判を上げることや競争者のレストラン評判を落とすことのような悪意を持ち、雇われた人たちによる不正レビュー・コメント情報を用いて、レストランの評判情報を任意的に操作する脅威が存在する。従来の手段では、数多くのレビュアーを雇うには莫大な費用がかかるだけではなく、既存の対策に検知される可能性もある。大規模な攻撃キャンペーンが実施できなくてインパクトが弱くなることは攻撃者が直面する問題である。また、最新の人工知能技術に関するソースコードはGitHubなどで公開されており、誰でも簡単且つ自由に入手できるようになっている。攻撃者の悪用に繋がりうる状況が整ってきた。そこで、本研究では、それを解決するために自然言語の自動生成技術の応用が最適であることを、攻撃者より先読みをして、最新の人工知能技術を悪用する攻撃手法の実現とその対策の確立を行うことを目指している。レストランの評価サイトにユーザが投稿する大規模かつ不均一なレビュー・コメント情報に深層学習をベースとした自然言語の生成手法を適用し、より人間の言葉に近い不正レビューを自動的かつ大量に創出することと、それらの不正レビューに対して既存の手法が対応できない高精度かつ高効率な検知アルゴリズムを開発することを狙いとした。




**アピールポイント**

本研究課題の創造性は、従来の研究では雇われた人たちが作成した不正レビューに着目する検出アルゴリズムが開発されてきた。これに対し、本研究は攻撃者の視点に立ち、自然言語生成技術を用いた大量の不正レビュー自動生成の実現とそれに相応しい解決方法の確立を目指すことにある。攻撃者は常に新しい攻撃手法を作り出し、不正レビューの検知を回避している。そのような戦略に対して、攻撃者の思考を先回りするセキュリティ対策が有効である。また，提案された攻撃手法により生成された不正レビュアーの検知方法を確立することによって，既存手法で検出できない領域がカバーされてレストランの評価サイトの健全性および信頼性の維持と更なる向上が期待できる。このような問題設定はレストランの評価サイトに限らず、ユーザが気軽にフィードバックを行うシステムに対しても有効であり、応用範囲は広い。