

AIシステムへの攻撃対策のためのアセット

株式会社KDDI総合研究所
2019年8月28日

機械生成文書の脅威

- AIの高度化により、様々な文書を自動生成可能
- 機械生成文書を悪用した攻撃が多様化



敵対的文書：
機械学習を用いた文書分類システムに対して、誤分類を誘発するために意図的に生成された文書

人間の文書 ポジティブと判定	<i>Almost Christmas is a movie that has balanced all its features to make a great movie.</i>
敵対的文書 ネガティブと誤判定	<i>Almost Christmas balances attributes to create a very very wonderful film.</i>

映画レビューの感情分析システムに対する敵対的文書の例

機械生成文書を検知するための特徴抽出モジュール

- 機械学習を用いた検知システムへの適用を想定
 - ・ 文書から特徴抽出を行うためのモジュールを提供
 - ・ 高精度な検知を可能とする独自開発の複数の特徴量へ変換



- 3種類の特徴抽出モジュールを提供
 - ・ 機械翻訳による盗作文書用
 - 逆翻訳による想定する文書の長さ：複数の文で構成された長めの文書
 - 対応言語：英語
 - ・ 偽造文書用
 - 想定する文書の長さ：1文から長文まで
 - 対応言語：英語
 - ・ 敵対的文書用
 - 適用可能な文書の長さ：1文から長文まで
 - 対応言語：英語

適用事例

● 機械翻訳による盗作文書

- 機械翻訳により独語から英語の文章を生成して盗作する攻撃を想定
- データセット：TEDのスクリプトを利用
- 文書の長さ：平均14.4文
- 検知精度：87%（従来手法 72.4%）

● 逆翻訳による偽造文書

- 仏語を経由した逆翻訳により英語文章を偽造する攻撃を想定
- データセット：European Parallel Corpus
- 文書の長さ：1文
- 検知精度：83.4%（従来手法 66.7%）

● 敵対的文書

- 映画レビューの感情分析システムに対する攻撃を想定
- データセット：Stanford Sentiment Treebank
- 文書の長さ：1文
- 検知精度：82%（従来手法 77%）