統計的フィルタリングに対するWord Salad攻撃についての考察

◎岩永学*，田端利宏**，櫻井幸一**

九州大学大学院 *システム情報科学府

**システム情報科学研究院

本研究の一部は，財団法人セコム科学技術振興財団平成15年度研究助成「インターネット妨害障害に対する暗号論的対策技術の研究」の支援を受けている．

発表の流れ

- 研究の背景
  - 迷惑メール
  - 統計的フィルタリング
  - 統計的フィルタリングへの攻撃：英語の場合
- 統計的フィルタリングへの攻撃：日本語の場合
  - 実験
- まとめ

迷惑メール

- 近年，急増
  - 1通あたりの送信コストが安価
- 迷惑メール送信の巧妙化
  - あからさまに迷惑メールとわかる単語や文章を避ける
  - 送信者のメールアドレスを詐称する
- 大量の迷惑メールにより生じる問題
  - ユーザの時間や通信コストの浪費，精神的苦痛
  - 重要なメールを見落とす危険性
  - ネットワークの構築･維持コストの増大
  - ➢ 電子メールシステムの危機

迷惑メール対策手法

- **ブラックリスト**
- **コンテンツの内容に基づくフィルタリング**
  - **ルールベース**
  - **統計的**
- **メールアドレス詐称の防止**
  - SenderID, SPF (Sender Policy Framework)
  - Challenge/Response
- **量的制限**
  - **一定量の作業を課す**(Computational Puzzle, Reverse Turing Test)
  - **応答の遅延**

etc...

統計的フィルタリング

- 過去のメールの特徴を元に，対象とするメールが迷惑メールかどうか判定する
  - 過去の正当なメールおよび迷惑メールから
  - 単語，文章，ルールなど
- 主な手法
  - Naive Bayes (Bayesian Filtering)
  - Boosting
  - Support Vector Machine (SVM)

  etc...

統計的フィルタリング

- **利点**
  - **おおむね正確(場合によっては99%以上)**
  - **新種の迷惑メールを学習可能**
  - **ユーザのPCにも導入可能**
  - **正当な送信者の協力が不要**

- **多くの導入例**
  - **メーラ**: Outlook, Mozilla, Eudora, …
  - **単体のフィルタ**:Popfile, SpamBayes, CRM114, …
    - **クライアントのPC上で動作**
    - **メールサーバ等の上で動作**

統計的フィルタリングへの攻撃

- 統計的フィルタリングの普及

↓

迷惑メール送信者による対策
(フィルタリングへの攻撃)

1. **統計的性質によらない攻撃**
   - **特徴(単語など)の抽出を妨害**
     - **単語を記号で分割**
     - **HTMLのタグや実体参照, エンコーディング**
     - JavaScript
     - **文章を画像として添付**

統計的フィルタリングへの攻撃(cont.)

2. **統計的性質に対する攻撃**
   - **無関係な単語等の挿入で確率計算に影響を与える**
     - "Word Salad攻撃"
     - **全くランダムな単語** (Dictionary Word Attack)
     - **迷惑メールに特徴的な単語を除く** (Common Word Attack)

J. Graham-Cumming, How to beat an Adaptive Spam Filter, 2004 Spam Conference

G. Wittel et al., On Attacking Statistical Spam Filters, CEAS 2004

Word Salad攻撃の例

From: spammer@example.com
To: some@recipient.com
Subject: University Diplomas, No Classes Needed
Date: Tue, 27 Jul 2004 13:52:39 -0500

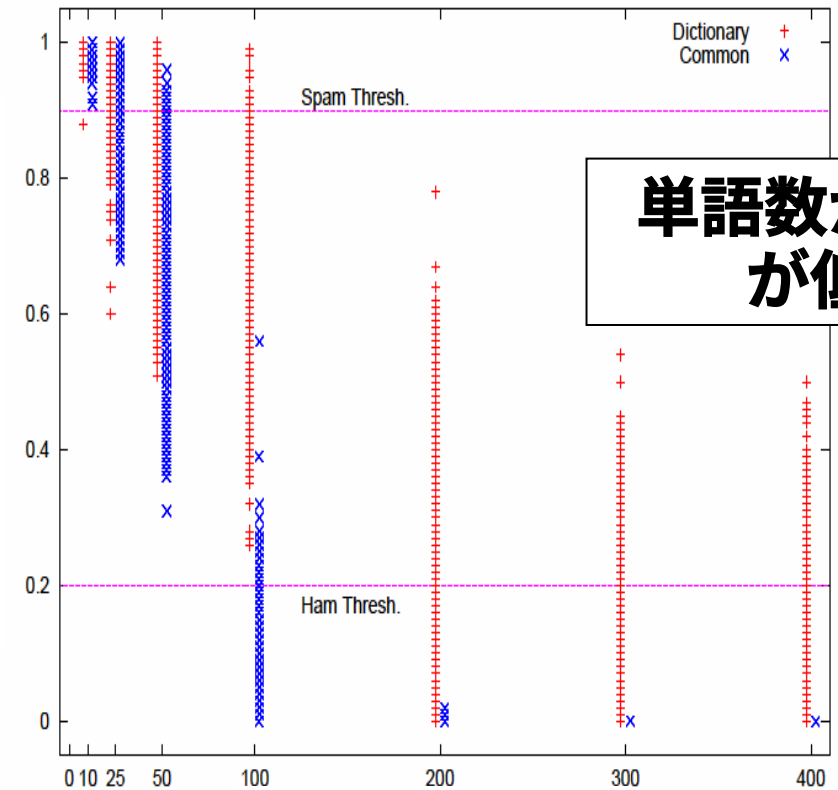
Academic-Qualifications from NON-ACCR. Universities.
No exams. No classes. No books.
Call to register and get yours in days - 1-234-567-8901

No more ads: removeme@example.com

pittsfield southern comeback memo egret droll rump pidgin promiscuous licentious farthest heater admiration b slothful egan afterimage skillful ingratitude necromancer innermost ftc curia zucchini molt dynamo panhandle topsoil daybed rheostat chaos upkeep augmentation sheer kaleidescope libelous curlicue beech persecute downstate cloudburst alewife decolonize hansom embarcadero destroy crumble serviceable fiske allegra deprivation bemoan sodden arteriolosclerosis bobcat caiman bladdernut aerodynamic satan requited bestowal midweek goodman iran debarring chard perception spanish ascription advisee cursor dibble bet rooky dragnet plant pragmatic

英語におけるWord Salad攻撃

- Naive Bayesに対する攻撃
  (Wittelらによる実験, *spamBayesを使用*)



迷惑メールと判定

単語数が増えるにつれ, 迷惑メール確率が低下(見逃されやすくなる)する

不確実(Unsure)と判定

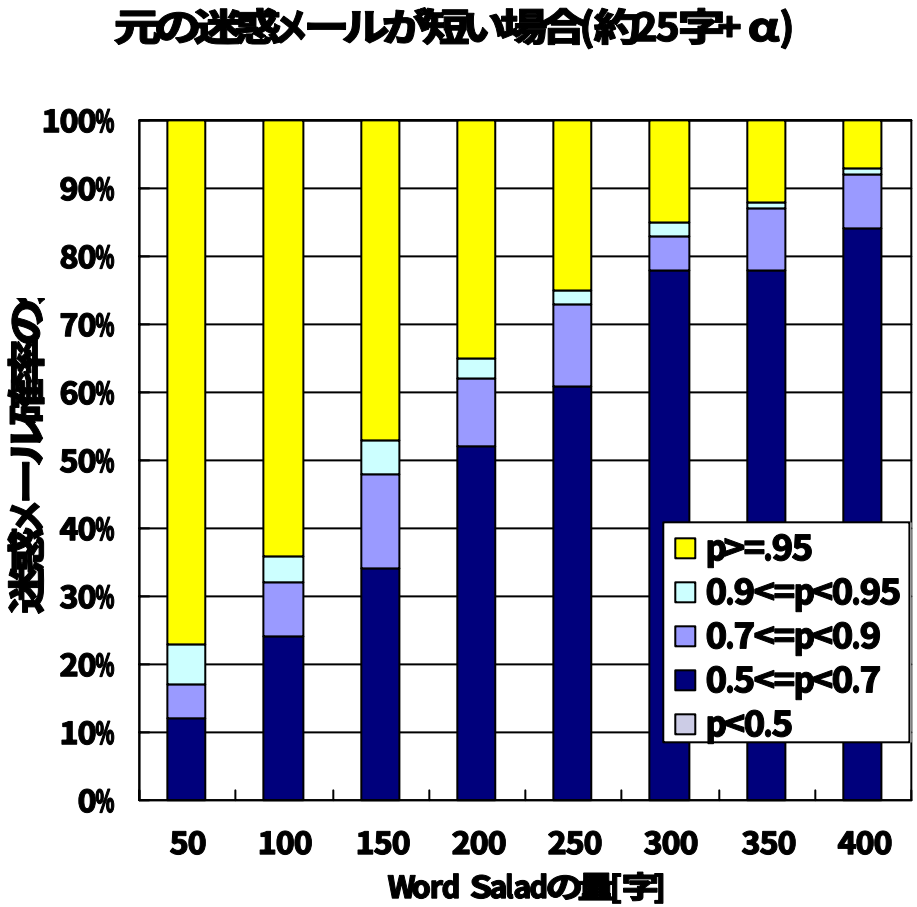
正当なメールと判定

G. Wittel et al., On Attacking Statistical Spam Filters, CEAS 2004

Word Salad攻撃：日本語の場合(実験)

- 迷惑メールに100通りのWord Saladを付加し、統計的フィルタリングの一種であるベイジアンフィルタに判定を行わせる
  - 迷惑メール本体が短い場合(約25字)
  - 迷惑メール本体が長い場合(約475字)
    - ベイジアンフィルタの実装のひとつであるBsfilterを使用
- 英語の場合との違い
  - トークンの抽出方法が画一的ではない
    - 文字の種類(漢字，ひらがな，カタカナ)を基準にトークンを抽出する方法
    - n個の連続する文字をトークンとして抽出する方法
    - 文法的に区切り，得られた単語をトークンとして抽出する方法
  - 付加する単位が単語ではトークンとして抽出されるとは限らない
  - 実験では単語ではなく文章を付加した
    - Web上のニュース記事をWord Saladとして使用

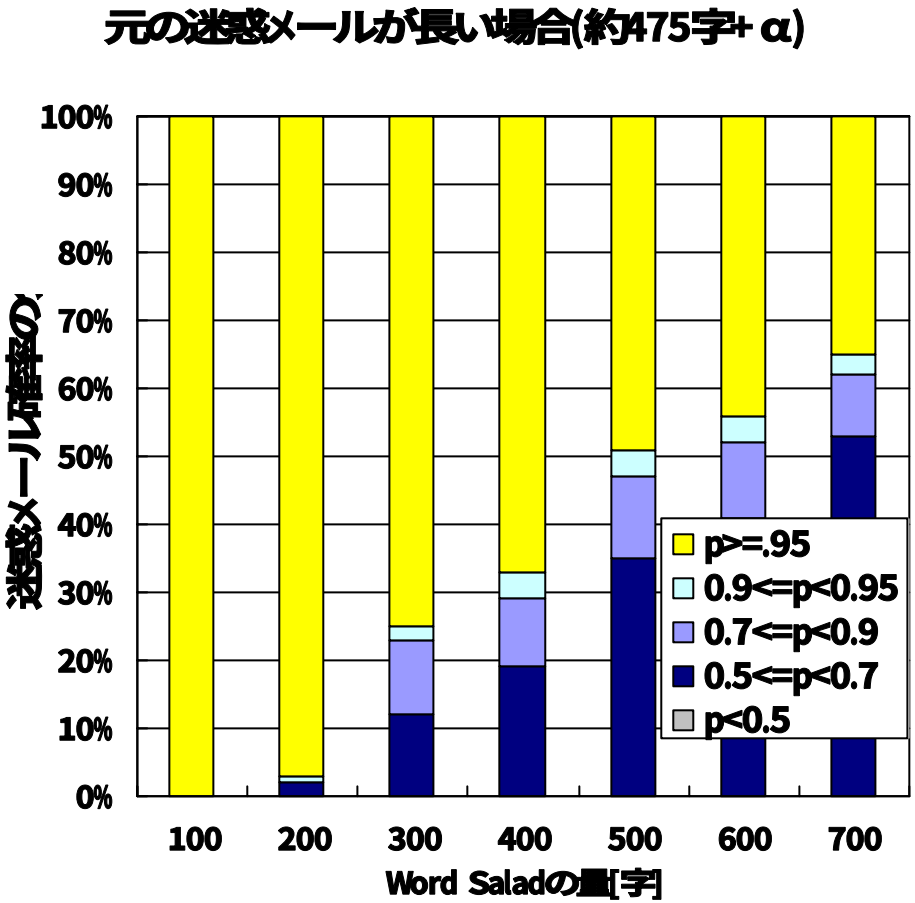
日本語の場合(結果)(1)



- Word Saladのない状態では，元の迷惑メールに対する迷惑メール確率は1.0

- **迷惑メールにWord Saladを付加することで，迷惑メール確率は低下する**
  - **＝迷惑メールと正当なメールの区別が難しくなる**
  - Word Saladの量が増えると，確率もより低下する

日本語の場合(結果)(2)



- Word Saladのない状態では，元の迷惑メールに対する迷惑メール確率は1.0

- **より長い迷惑メールに対しては，相対的にWord Saladの効果は小さくなる**
  - **より多くのWord Saladが必要**

まとめ

- **統計的メールフィルタリングへの攻撃**
  - **統計的性質によらないもの**
  - **統計的性質に対するもの**: Word Salad
  - Word Salad: **日本語の場合**
- Word Salad**は日本語の迷惑メールにおいても効果がある**