情報処理学会論文誌 Vol.54 No.9 2156–2166 (Sep. 2013)

推薦論文

マルコフ連鎖による合成文章の不自然さを用いたCAPTCHAの提案と安全性評価

鴨志田 芳典[1,a)] 菊池 浩明[2,b)]

受付日 2012年12月1日, 採録日 2013年6月14日

概要：ボットによるアカウントの大量取得や，それにともなう不正行為への対策として広く用いられているCAPTCHAと呼ばれる機械判別方式は，コンピュータには判別が困難だが人間には容易である問題を利用することでプログラムによる入力と人による入力とを識別する．本稿ではOCR機能を持ったマルウェアやリレーアタック（クラウドソーシング）に耐性を持つCAPTCHAとして，ワードサラダと呼ばれるマルコフ連鎖による文章合成の不自然さを用いたCAPTCHAを提案する．提案手法への先にあげた攻撃と，文章校正を用いた攻撃に対する安全性の評価を行い，その改良方式を示す．さらに，日本語以外へ適用した実験結果を示す．

キーワード：CAPTCHA，アクセス制御・認証，マルコフ連鎖，文章合成

Proposal of CAPTCHA Using Artificial Synthesis Sentences and Its Security Evaluation

Yoshifumi Kamoshida[1,a)] Hiroaki Kikuchi[2,b)]

Received: December 1, 2012, Accepted: June 14, 2013

Abstract: The CAPTCHA is a widely used technique to prevent malicious software, called bot, from obtaining false account. It uses an easy problem for human but difficult for machines to distinguish an input made by a program and one by human. We have proposed a new CAPTCHA testing sentences synthesized from the Markov chain model, called “word salad”, which is tolerant against malware with an OCR feature, or a “crowd sourcing” attack, called as “relay attack”. In this paper, we estimate the security against the attack using proof-reading tool to distinguish the synthesized sentences, and present an improved protocol for the attack. Moreover, we show an experimental result of some languages other than Japanese.

Keywords: CAPTCHA, authentication, Marcov chain model, synthesize sentences

1. はじめに

近年，ボットによるアカウントの大量取得やそれにともなう不正行為への対策としてCAPTCHAと呼ばれる不正プログラム判別方式が広く用いられている [1]．CAPTCHAはコンピュータには判別が困難だが人間には容易である問題を利用することで，ボットやエージェント等のプログラムされた入力と人による入力とを識別する．しかしながら，最も広く用いられている文字列画像を変形させたCAPTCHAは，高精度のOCR機能を持ったマルウェアによって破られてしまうことが報告されている [2], [4]．人間を使ったリレーアタック（クラウドソーシング）と呼ばれる攻撃も問題になっている [7]．リレーアタックでは，攻撃者はCAPTCHAの問題を自分の運営するWebサイトに転

1 東海大学大学院工学研究科
Graduate School of Engineering, Tokai University, Hiratsuka, Kanagawa 259–1292, Japan
2 明治大学総合数理学部
School of Interdisciplinary Mathematical Sciences, Meiji University, Nakano, Tokyo 164–8525, Japan
a) y.kamoshida2@gmail.com
b) kikn@meiji.ac.jp

本稿の初期の版は第26回ファジィシステムシンポジウム（FSS2010）と，マルチメディア，分散，協調とモバイルシンポジウム（DICOMO2012）にて発表している．

載し，それを人に解かせることにより CAPTCHA を成功させる．転載された CAPTCHA を解く人間は発展途上国の低賃金労働者である．したがって CAPTCHA には，(1) 問題の合成が機械的に可能であること，(2) 低賃金労働者による攻撃に頑強であること，(3) 不正プログラムには判別が困難であること，という必要条件がある．

そこで，視覚的な情報だけではなく，人間のより高度な認知処理を用いた CAPTCHA の研究が行われている．Assira [3] はコンピュータが画像の意味を理解することの困難さを利用した代表的な CAPTCHA で，画面上に表示された複数の画像から，犬の画像か猫の画像のみを選択させることにより人間と不正プログラムを区別する．山本らは繰り返し機械翻訳された文章の不自然さを用いた CAPTCHA を提案している [5], [6]．鈴木らはユーザ側の Web ブラウザにアドオンを追加することでリレーアタックを防ぐ方式を提案している [7]．

我々の研究もこの試みの 1 つであり，リレーアタックに耐性を持つ CAPTCHA として，ワードサラダと呼ばれるマルコフ連鎖よる文章合成の不自然さを用いた CAPTCHA を提案する．ワードサラダはスパムメールやスパムブログの大量投稿に用いられる手法であり，Web から収集した文章から作成した n-gram 頻度データを基に n 階マルコフ連鎖により確率的に文を合成する．ワードサラダはコーパスの特徴を反映した文法的に正しい文章を合成するが，合成された文章は人が見れば話題のつながり方等から不自然であると容易に判断できる．文法的には正しいため，コンピュータには判別が困難である．提案手法では自然な文とする文章とワードサラダを合成するためのコーパスとなる文章を収集し，コーパスから合成したワードサラダと自然な文をランダムに 1 つずつ提示し，設定された閾値以上の精度で正しく答えられるか否かで人と不正プログラムを判別する．

ワードサラダは，1 つのコーパスから異なる文を大量に合成できるため，条件 (1) を満たしている．一般的なアルゴリズムであるので，日本語以外にも適用可能である．提案手法では文の不自然さを理解できるネイティブレベルの言語能力が必要となるため，外国人低賃金労働者を用いたリレーアタックに対しても耐性があり，条件 (2) も満たしていると期待される．

しかしながら条件 (3) の不正プログラムには，単純な総当り攻撃だけでなく自然言語の知識ベースによる文章校正機能による攻撃も考えられ，その効果は自明ではない．たとえば，**表 1** のように，文法や用法の誤りならば機械的に修正できるからである．

そこで本稿では，ワードサラダが Microsoft Word の文章校正ツールによって校正されるときに，不正プログラムによって判別できると仮定し，その安全性を厳密に評価する．日本語以外に英語，中国語での提案手法の実装を行い，

表 1 ワードサラダに対する文章校正の例
Table 1 Samples of corrections for synthesized sentences, "words salada"

1．校正が行われた
第二次世界における影響力は、各国の影響力を樹立して いったの クリストファー・検閲等から遠洋捕鯨が 民間に送られてさらに各地から購入した。
2．校正が行われない
バラク・オバマ大統領の紛争や国民に対して 政治的に殆ど被害を謳歌している。

その有効性から提案手法の他言語への適用可能性について検討する．以上の評価結果に基づき，攻撃者が一定の確率で問題が検知できるときに，それを考慮した場合の提案手法の最適な閾値と，それについての精度とパフォーマンスについて評価する．

本稿の構成は次に示すとおりである．まず，2 章でマルコフ連鎖による文章合成のアルゴリズムと，文章校正の概要を示す．3 章では文章合成の不自然さを用いた CAPTCHA と，それに対する攻撃モデルを提案する．4 章では提案手法の安全性を評価する．5 章ではコーパスを用いて 5 つの実験を行い，提案方式の精度と安全性を評価する．6 章で本研究を結論づける．

2. 要素技術

2.1 n 階マルコフ連鎖による文章合成のアルゴリズム

n 階マルコフ連鎖 [8] による文章合成は，コーパスから抽出した n-gram 頻度データに基づいてマルコフ連鎖モデルを作り，人工的な文章を合成する手法である．マルコフ連鎖による文章合成で i 番目に出力される語 x_i は条件付き確率

$$P(x_i) = P(x_i|x_{i-1}, x_{i-2}, \ldots, x_{i-n})$$

に従う．日本語の場合は分かち書きされていないため，単語単位の n-gram 頻度を得るために前処理として形態素解析器による分かち書きが必要となる．抽出した単語 n-gram 頻度データを基に n 階マルコフ連鎖で文章を合成する．以降，マルコフ連鎖により合成された文章をワードサラダと表記し，ワードサラダ合成に用いたマルコフ連鎖の階数を階数 n とする．前述のように n-gram 言語モデルでは，n-gram で切り出された語集合での $n-1$ 番目の語から n 番目の語を推測することが可能である．そのため，n 階マルコフ連鎖で文章を作るためには $(n+1)$-gram 頻度データが必要となる．

2.2 文章校正

文章校正機能とは，文章として間違っている箇所を自然

本稿の内容は 2012 年 7 月のマルチメディア，分散，協調とモバイル（DICOMO2012）シンポジウム 2012 にて報告され，コンピュータセキュリティ研究会主査により情報処理学会論文誌ジャーナルへの掲載が推薦された論文である．

言語処理により検出・修正する機能である．一般に広く使われているワープロソフトウェアにも搭載されており，手軽に利用できる．$n = 1$ のワードサラダにおいて，Microsoft Word 2007 の文章校正機能により校正が行われた文と行われなかった例をそれぞれ表 1 に示す．文 1 では，入力ミスと判断され校正が行われた．文 2 はこのように，文の意味は支離滅裂であるが校正は行われていない．文章校正機能はある程度スパム文書を検出する機能を有する．

3. 提案方式

3.1 原理

本稿で提案する文章合成の不自然さを用いた CAPTCHA では，コーパスから合成したワードサラダと自然な文をランダムに 1 つずつ提示し，設定された閾値以上の精度で「自然」か「不自然」かを正しく答えられるかどうかで人と機械を判別する．この方式はワードサラダの判別が人間にとってはやさしく，機械にとって困難であることを仮定している．

3.2 提案方式

提案方式は，コーパスの収集，マルコフ連鎖モデルの作成，問題となる文の合成，CAPTCHA による認証からなる．

3.3 コーパスの収集

文章合成のためのコーパスは Web をクローリングし，新聞社の最新ニュース記事等を利用して収集する．ただし，Web から検索できる文を自然な文にした問題では，それを Web で検索することで容易に判断可能である．よって，新聞のアーカイブ等の有料コンテンツ[*1]や，青空文庫[*2]のように，無料ではあるが直接 Web からは検索できない文章をコーパスとすることが望ましい．

3.4 マルコフ連鎖モデルの作成

マルコフ連鎖モデルは，単語の頻度データを収集したコーパスから作成する．提案手法ではより不自然な文章を合成する方が CAPTCHA 精度は高くなることが予想されるため，最も不自然な文を合成する確率の高い階数での文章合成を行う．文章の不自然さという点では単語をランダムに結んだ文章の方がより不自然ではあるが，文法解析により容易に検出されてしまう．一方，ワードサラダは品詞の並びが文法として適切となるという特徴がある．

3.5 CAPTCHA による認証

コーパスから自然な文を h 個とマルコフ連鎖で合成したワードサラダ s 個を合わせて $c = h + s$ 個の文の集合を用意する．その際，合成された文章の内コーパスの一部と完全に一致する文は除外する．自然な文とワードサラダを h/c と s/c の確率でランダムに提示し，ユーザ対して「自然」,「不自然」の選択を求める．自然な文に対して「自然」と正答した回数と，ワードサラダに対して「不自然」と正答した回数の合計を正解数 k とし，k があらかじめ定めた閾値 θ 以上ならば CAPTCHA 成功（人間である）とする．提案方式をアルゴリズム 1 に示す．

Algorithm 1 提案方式

1 コーパスから n 階マルコフ連鎖モデルを作る．
2 自然な文を h 個，ワードサラダを s 個，計 c 個の文をランダムな順でユーザに与える．
3 ユーザは c 個の文を「自然 H」と「不自然 S」のどちらかに分類して回答する．
4 正答数 k を求め，$k \geq$ 閾値 θ ならばユーザを受理，$k < \theta$ ならば拒否する．

3.6 攻撃モデル

3.6.1 総当り攻撃（Brute force）

提案方式において最も容易なボットによる攻撃手段として，すべての問いに対しランダムで解答する総当り攻撃が考えられる．h と s の比率を知らない攻撃者は 1/2 の確率でランダムに解答する．5.8 節ではその比率を知る攻撃者も考える．

3.6.2 人による攻撃（リレーアタック）

リレーアタックでは，攻撃者は CAPTCHA の問いを自分の運営する Web サイトに転載し，それを人に解かせることにより CAPTCHA を成功させる．転載された CAPTCHA を解く人間は標的サイトを母国としない発展途上国の低賃金労働者である．

3.6.3 自然言語処理の知識を用いた攻撃（ワードアタック）

ワードサラダは機械による自動的な判別が困難であるとされているが，自然言語処理により不自然な文章を合成する以上，同様に自然言語処理を用いることで攻撃の精度を高めることが可能であると予測される．提案方式への自然言語処理を用いた攻撃としては，スパムらしい文章を検出するベイジアンフィルタ [9], [10] のほか，ワードサラダの特徴に着目して検出する手法 [11], [12] や，文章として間違っている箇所を検出・修正する文章校正機能が考えられる．

4. 安全性評価

4.1 安全性の定義

X を出題文を表す確率変数，Y を解答文を表す確率変数，H を人間による文章，S をスパム（機械生成の）文章とすると，自然な文を出題して自然と回答する条件付き確率は $P(Y = H|X = H)$ と表せる．自然な文章とスパム文章を出題する確率はそれぞれ，

[*1] 朝日新聞，http://www.asahi.com/information/db/forb.html
[*2] 青空文庫，http://www.aozora.gr.jp

$$P(X=H)=\frac{h}{c}$$
$$P(X=S)=\frac{s}{c}=1-\frac{h}{c}$$

である．したがって，自然な文章とスパム文章の出題数の比率を考慮したCAPTCHA成功率は，これらの同時確率で

$$P(Y=H,X=H)=P(Y=H|X=H)P(X=H)$$
$$P(Y=S,X=H)=P(Y=S|X=H)P(X=H)$$
$$P(Y=H,X=S)=P(Y=H|X=S)P(X=S)$$
$$P(Y=S,X=S)=P(Y=S|X=S)P(X=S)$$

と与える．CAPTCHA検査の失敗には，正しい自然な文章をスパムと誤判定することと，スパム文章を自然な文章と誤判定することの2種類があり，これらをまとめて，人間によるCAPTCHA失敗率

$$P_q=P(Y=S,X=H)+P(Y=H,X=S)$$

とする．

ここで，人間がCAPTCHAを試行したのに$k<\theta$となる確率を，人間拒否率FRR（False human Rejection Rate）と定める．

次に機械によるCAPTCHA成功率をP_mとする．機械による攻撃が$k\geq\theta$を満たす確率を機械受入率FAR（False machine Acception Rate）と定める．

このとき，FRRおよびFARは確率P_qおよびP_mの二項分布で

$$FRR=\sum_{k=\theta}^{c}\binom{c}{k}{P_q}^k(1-P_q)^{c-k}$$
$$=\sum_{k=0}^{\theta-1}\binom{c}{k}(1-P_q)^{c-k}{P_q}^k$$
$$FAR=\sum_{k=\theta}^{c}\binom{c}{k}{P_m}^k(1-P_m)^{c-k}$$

と表すことができる．また，$FRR=FAR$となる値をEER（Equal Error Rate）とする．本稿ではこれらの値をCAPTCHAの安全性の精度とする．

4.2 総当り攻撃に対する安全性

3.4.1項で述べた総当り攻撃のうち，hとsの比率を知らない攻撃者の場合，1つの問いに対して正解する確率P_{mb}は1/2であり，総当り攻撃でk回正解する確率FARは$\frac{1}{2^k}\sum_{k=\theta}^{c}\binom{c}{k}$となる．たとえば，$c=h+s=15$，$\theta=13$のとき，$FAR$は約0.37%である．

4.3 ワードアタックに対する安全性

ここでは，ワードアタックに対する評価方式を考える．1つの文について文章校正を表す確率変数をWとする．すなわち，校正が行われる事象を$W=t$とし，その確率を$P(W=t)$とする．ここで，$P(W=t)$は

$$\begin{aligned}P(W=t)&=P(W=t,X=S)+P(W=t,X=H)\\&=P(W=t|X=S)P(X=S)\\&\quad+P(W=t|X=H)P(X=H)\end{aligned}\tag{1}$$

となる．このとき，文章校正が行われた文の入力がスパムである確率$P(X=S|W=t)$は，ベイズの定理により

$$P(X=S|W=t)=\frac{P(W=t|X=S)P(X=S)}{P(W=t)}\tag{2}$$

で与えられる．同様に校正が行われなかった事象を$W=t$とし，その確率を$P(W=f)$とする．ここで，$P(W=f)$は

$$\begin{aligned}P(W=f)&=P(W=f,X=S)+P(W=f,X=H)\\&=P(W=f|X=S)P(X=S)\\&\quad+P(W=f|X=H)P(X=H)\end{aligned}\tag{3}$$

となる．このとき，校正が行われなかった際の入力が自然な文である確率は

$$P(X=H|W=f)=\frac{P(W=f|X=H)P(X=H)}{P(W=f)}\tag{4}$$

である．したがって，機械の解答Y_wを，$W=t$のとき，

$$Y_w=\begin{cases}S & \text{w./p. } P(X=S|W=t)\\ H & \text{w./p. } P(X=H|W=t)\end{cases}\tag{5}$$

$W=f$のとき

$$Y_w=\begin{cases}S & \text{w./p. } P(X=S|W=f)\\ H & \text{w./p. } P(X=H|W=f)\end{cases}\tag{6}$$

とそれぞれ定めることで，機械受け入れ率を最大化できる．

この判定機による判定が実際に成功するのは，$P(Y_w=S,X=S)$と$P(Y_w=H,X=H)$の2種類である．すなわち，この場合の機械の1題あたりのCAPTCHA成功率は，

$$P_{mw}=P(Y_w=S,X=S)+P(Y_w=H,X=H)\tag{7}$$

となる．ここで，

$$\begin{aligned}&P(Y_w=S,X=S)\\&=P(Y_w=S|W=t)P(W=t|X=S)\\&\quad+P(Y_w=S|W=f)P(W=f|X=S)\end{aligned}\tag{8}$$

$$\begin{aligned}&P(Y_w=H,X=H)\\&=P(Y_w=H|W=t)P(W=t|X=H)\\&\quad+P(Y_w=H|W=f)P(W=f|X=H)\end{aligned}\tag{9}$$

である．

5. 実験

提案方式の有効性を検証するため，次の項目をあげる．

5.1 実験項目

実験 1 人間がワードサラダを不自然と判断するできるマルコフモデルの階数はどのくらいか？

実験 2 自然な文を正しく判別する確率とワードサラダを正しく検出する確率の 2 種類の正答率はいくらか？

実験 3 文章校正に対する安全性はどの程度か？

実験 4 リレーアタックに対する安全性はどの程度か？

実験 5 提案方式は日本語以外も同様に適用できるか？

5.2 実験方法

5.2.1 評価データとマルコフモデル

評価データは，不自然な文とする階数 $n = 1,\ldots,5$ のワードサラダと，自然な文とする合成に用いた文章の一部切り抜きである．実験に用いるマルコフモデルは 5,000 文字程度の政治・経済に関する記事から学習する．

5.2.2 実験 1（マルコフモデルの階数の効果）

情報系の学生 8 名を被験者とする．5 文からなる評価データ（$s = 50$，$h = 50$ の計 100 題）を順次提示し，その文章が機械的に合成されたものであるかどうかを判断させ，その正答率と解答にかかる時間を計測する．

5.2.3 実験 2（正答率）

7 名の被験者に対し，100 題ずつ用意された 1 文からなる評価データを $h = 5$，$s = 10$ の比率でランダムに 15 回提示する．文が自然か不自然かを判断させ，自然な文を正しく自然と判定する確率 $P(Y = H|X = H)$ とワードサラダを正しく不自然と判定する確率 $P(Y = S|X = S)$ の 2 種類の正答率と解答にかかる時間を計測する．実験は 1 人につき複数回行い平均を求める．

5.2.4 実験 3（ワードアタックの効果）

ワードサラダ 300 文と自然な文を 300 文を出力し，Microsoft Word 2007 を用いて文章校正をする．文中に校正箇所が示された文を検出が行われたと見なして，その確率 $P(W = t|X = S)$ を計測する．

5.2.5 実験 4（リレーアタックの効果）

日本語を学んだ留学生をリレーアタックの低賃金労働者と見なし，その攻撃成功率を調べる．提案手法の CAPTCHA を解くためにある程度日本語を学習した攻撃者が存在する場合を想定し，日本語を学んだ留学生 3 名を被験者とする．実験に用いたマルコフモデルは，青空文庫から収集した 4 つのテキストから抜き出した 20,000 文字程度のコーパスから学習する．

5.2.6 実験 5（他言語適用性）

日本人学生 3 名，英語，中国語を母国語とする学生それぞれ 1 名に対し各言語の評価データを提示し，正答率を計測する．評価データは，文章の意味を揃えるために Wikipedia[*3] のアメリカ合衆国の記事の本文から合成する．ワードサラダでは，括弧が単語と見なされて対応しない括弧が生成されてしまうので，括弧表記（「」，()，＜＞，≪≫等）はすべて削除した．日本語と中国語の形態素解析器にはそれぞれ，MeCab [13] と ICTCLAS [14] を用いる．英語はスペースのみで分かち書きを行う．実験で使用した問題を付録に示す．

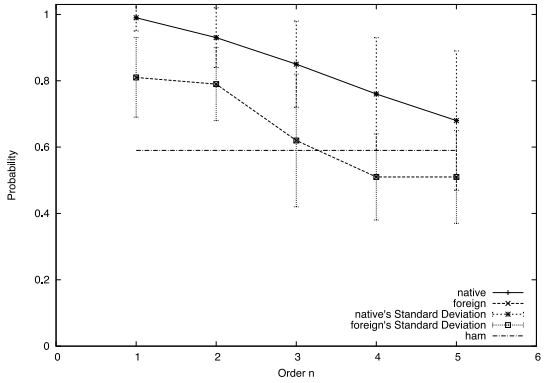

図 1 階数 n についてのスパム検出率 $P(Y = S|X = S)$ と自然な文に対する正答率 $P(Y = H|X = H)$

Fig. 1 Probabilities of correct answers with respect to order n, $P(Y = S|X = S)$ (detection rate of Spam) and $P(Y = X|X = H)$ (accuracy of natural sentence), respectively.

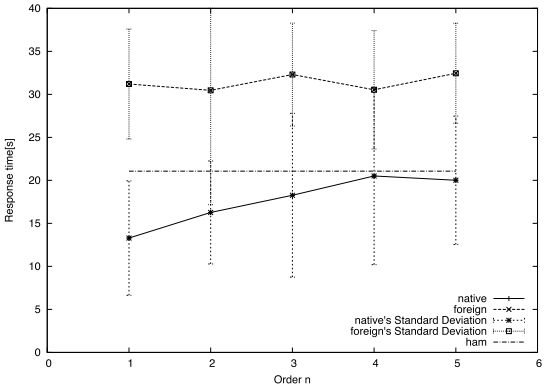

図 2 階数 n についての応答時間 [s] と自然な文に対する応答時間 [s]

Fig. 2 Response time with respect to order n for synthesized and natural sentences.

5.3 実験結果

実験 1，実験 4 でのワードサラダに対する正答率 $P(Y = S|X = S)$，$P(Y = H|X = H)$ と平均応答時間をそれぞれ**図 1** と**図 2** に示す．図 1 と図 2 ではそれぞれ，階数 n を Order n と表す．凡例の “native” と “foreign” は日本人学生と留学生をそれぞれ表し，“Standard Deviation” はそれらの標準偏差を示す．“ham” は自然な

[*3] Wikipedia, http://ja.wikipedia.org

表 2 正答率 $P(Y=S|X=S)$ と $P(Y=H|X=H)$
Table 2 Probabilities of correct answers $P(Y=S|X=S)$ and $P(Y=H|X=H)$.

出題	$n=1$	$n=2$	$n=3$
$X=H$	0.91	0.80	0.68
$X=S$	0.73	0.62	0.45

表 3 応答時間 [秒]
Table 3 Response time [second].

出題	$n=1$	$n=2$	$n=3$
$X=H$	8.05	8.12	7.44
$X=S$	6.19	7.75	8.58

表 4 条件付確率 $P(Y|X)$（文章量が 1 文，階数 $n=1$）
Table 4 Conditional probabilities of answer Y given input X(one sentence, order $n=1$).

出題文書 \ 判別文書	$Y=H$	$Y=S$
$X=H$	0.91	0.09
$X=S$	0.27	0.73

表 5 文章校正が行われる確率 $P(W=t|X=S)$
Table 5 Conditional probabilities of sentence to be corrected given $X=S$, $P(W=t|X=S)$.

	$n=1$	$n=2$	$n=3$	自然な文
$P(W=t|X=S)$	0.24	0	0	0

表 6 各言語の不自然な文の判別精度 $P(Y=S|X=S)$
Table 6 Detection rates of synthesized sentence for some languages.

言語	$n=1$	$n=2$	$n=3$	自然な文
日本語	0.87	0.47	0.20	0.90
英語	1.0	0.8	0.6	0.7
中国語	1.0	0.8	0.5	0.7

文に対する結果である．これらの結果から正答率，応答時間ともに最小となるのは $n=1$ で合成したワードサラダである．

実験 2 の正答率 $P(Y=S|X=S)$ と $P(Y=H|X=H)$ を**表 2** に，応答時間を**表 3** にそれぞれ示す．以上の結果より，階数 $n=1$ の正解率と誤り率を**表 4** の条件付き確率で整理する．

実験 3 で得られた文章校正が行われた割合を**表 5** に示す．実験 3 では $n=1$ のときのみ，300 件中 72 件，すなわち 24%の割合で文章校正が行われた．$n=1$ のワードサラダにおいて，校正が行われた文と行われなかった例をそれぞれ表 1 に示した．

実験 5 の結果を**表 6** に示す．各言語とも階数 n の値が低いときにスパム文書に対し不自然であると感じる割合は高くなる．

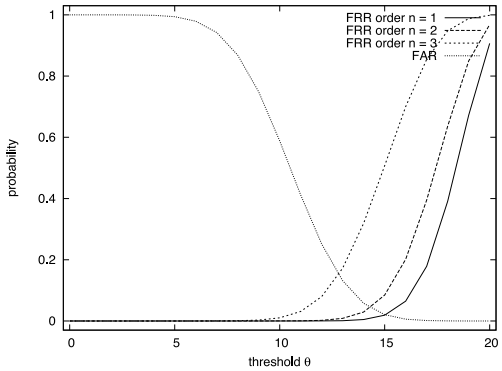

図 3 総当り攻撃に対する安全性（閾値 θ についての FRR と FAR の分布（$s=5$，$h=15$））
Fig. 3 Security for brute-force attack Distributions of FRR and FAR with respect to theta against brute-force attack.

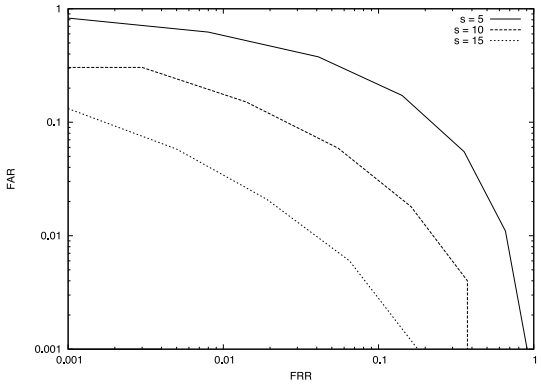

図 4 総当り攻撃に対する ROC 曲線
Fig. 4 ROC curve against brute-force attack.

5.4 総当り攻撃モデルに対する安全性

総当り攻撃では，攻撃者の 1 題あたりの成功率は $P_m=1/2$ である．階数 $n=1,\ldots,3$ のワードサラダにおいて $h=5$，$s=15$ の場合，閾値 θ についての FRR と FAR を**図 3** に示す．図 3 より $n=1$，$h=15$，$s=5$ の場合，最も EER に近くなる θ の値は $\theta=15$ のときである．s の値を変化させたときの FRR と FAR の関係を**図 4** に示す．

したがって，この条件で文章合成の不自然さを用いた CAPTCHA の FRR と EER はともに約 3%である．実験 2 より，自然な文章に対しての応答時間は 8.05 秒であり，合成された文章では 6.19 秒である．よって，1 回の CAPTCHA にかかる平均時間は 151.7 秒である．

5.5 ワードアタックに対する安全性

実験 3 より，Microsoft Word 2007 による文章校正が行われる確率は

$$P(W=t|X=S)=0.24$$
$$P(W=t|X=H)=0$$

と与えられた．$s=15$，$h=5$ で提案手法を適用した場

表 7 スパム判定機による出力 Y_w の条件付き確率 $P(Y_w|X)$
Table 7 Conditional probabilities of Spam detection machine Y_w given input X.

入力文書 \ 判別文書	$Y_w = H$	$Y_w = S$
$X = H$	0.798	0.202
$X = S$	0.394	0.606

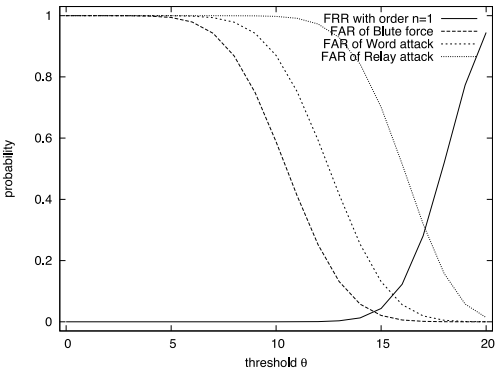

図 5 各攻撃手法の FAR
Fig. 5 Distributions of FAR for several attacks.

合，自然な文 H が出題される確率は $P(X = H) = 0.75$，不自然な文 S が出題される確率は $P(X = S) = 0.25$ となる．よって式 (1) より，提案手法において 1 題あたりに校正が行われる確率は $P(W = t) = 0.06$，行われない確率は $P(W = f) = 0.94$ となる．このとき，スパム判定機を用いた攻撃では，機械は文章校正による検出が行われれば必ず「不自然」であると解答し，そうでない場合は式 (4) より $P(X = H|W = f) = 0.60$ の確率で「自然」を，$P(X = S|W = f) = 0.40$ の確率で「不自然」を選択する．この「自然」「不自然」の選択は機械による出力がそれぞれ $Y_w = H$，$Y_w = S$ となることを表す．これらを整理すると，機械による攻撃の成功率 P_{mw} は**表 7** の確率 $P(Y_w|X)$，すなわち $P_{mw} = 0.697$ となる．

5.6 リレーアタックに対する安全性

リレーアタックによる効果については，実験 4 で得られた留学生による実験の結果から評価する．留学生の自然な文に対する正答率を $P(Y = H|X = H) = 0.5$ と仮定し，留学生による 1 題あたりの CAPTCHA 成功率は $P_{mc} = 0.725$ とする．そこから導き出される FAR をリレーアタックの精度と仮定する．被験者は日本語を数年間学んだ者であり，標的サイトを母国としない人間よりかなり高い精度であることに注意されたい．

5.7 3 つの攻撃モデルに対する安全性の比較

提案方式に対する，$n = 1$，$s = 5$，$h = 15$，$c = 20$ の場合において，閾値 θ についての 3 つの攻撃モデルの FAR と FRR を**図 5** に，ROC 曲線を**図 6** にそれぞれ示す．

以上の評価より，機械により文章校正ツールを用いた

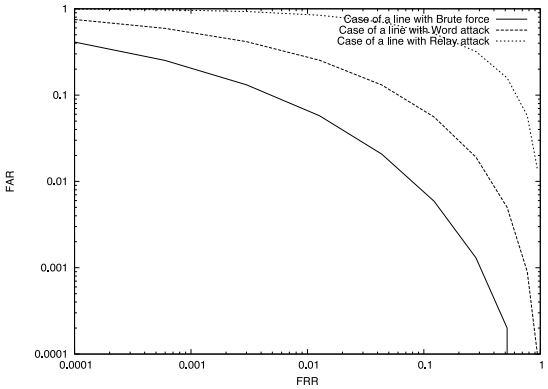

図 6 各攻撃手法の ROC 曲線
Fig. 6 ROC curves for several attacks.

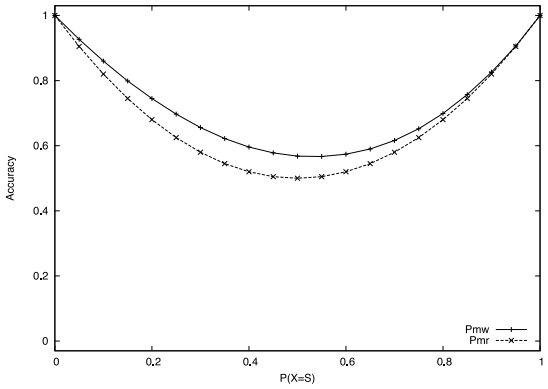

図 7 $P(X = S)$ についての P_{mw} と P_{mr}
Fig. 7 Improvement of accuracy given known probability of Spam $P(X = S)$.

ワードアタックが行われたとき，EER は 23%（閾値 $\theta = 17$ のとき $FRR = 24\%$，$FAR = 22\%$）となる．リレーアタックについては，$\theta = 18$ のとき $FRR = 36\%$，$FAR = 34\%$ となる．

5.8 最適なスパム出題率

5.7 節では，自然な文と不自然な文の出題数の偏りも精度の低下を招く大きな要因となった．そのため，CAPTCHA として提示する自然な文と不自然な文の出題数は偏らないことが望ましい．$c = 20$ のとき，s と h の値の割合を攻撃者が知っていれば，総当り攻撃よりも高い確率で受理される．この知識を持つ攻撃者による CAPTCHA 成功率を

$$\begin{aligned} P_{mr} &= P(Y = S, X = S) + P(Y = H, X = H) \\ &= P(Y = S)P(X = S) + P(Y = H)P(X = H) \\ &= \left(\frac{s}{c}\right)^2 + \left(\frac{h}{c}\right)^2 \end{aligned}$$

とする．一方，ワードアタックを用いたときの成功率は式 (7) の P_{mw} である．これらの攻撃の成功率 P_{mw} と P_{mr} を $P(X = S)$ $(= s/c)$ について比較すると，**図 7** のように分布する．図 7 より，P_w が最小になるのは s と h が同数の場合ではなく，$P(X = S) = 0.55$，すなわち，s と h の割

合が 4.5 : 5.5 のときであることが分かる．

ただし，人間の 1 題あたりの CAPTCHA 失敗率 P_q も問題に含まれるスパムの割合により変化するため，EER の最低値はそれほど自明ではない．たとえば $c = 20$ の場合では，$s = 9$ のときに $EER = 0.15$ となり，最小化する．

5.9 まとめと考察

実験 1，実験 2 から，日本人と留学生の間に日本語の自然な文を判別する能力に差があることが分かる．階数 n に依存して正答率が減少することからワードサラダは階数を増やす程自然な文と判別がつきにくくなる．図 5 より，階数 $n = 1$ のとき，最も EER に近くなる θ の値は $\theta = 15$ である．

実験 2 の応答時間の結果から，平均時間は 151.7 秒である．従来の CAPTCHA の解答にかかる応答時間は約 8 秒程度であるため，ユーザにかかる負担はとても高いといえる．

安全性に対する評価では，h と s の比率を知らない総当り攻撃を想定した場合と比較して，文章校正ツールを利用した攻撃の場合の EER は約 12%増加し 15%となった．ワードサラダの検出を Microsoft Word 2007 により行ったが，ワードサラダのフィルタリングに用いる手法を利用した検出も可能である．

実験 5 の結果では，日本語，中国語，英語については各言語とも同じような振舞いをするという結果が得られた．このほかにタイ語でも同様の実験を行ったが，形態素解析が適切に行えず信憑性のある結果が得られなかった[*4]．適切な形態素解析さえ行えれば提案手法は他言語にも適用可能であると考えられる．

5.10 自然な文に対する判別精度についての考察

実験から得られた，人間の自然な文に対する正答率 $P(Y = H|X = H)$ は 7 割から 9 割程度となっている．これらの値は直感的に，自然な文を自然と判断できる確率としては低い．この理由については，以下の点が考えられる．

1. **不注意な被験者の影響** 被験者から得られたデータのうちには，回答時間が極端に短く，提示された文書をよく読まずに直感的に回答を行ったと思われるものもあった．これらは提案方式のユーザにかかる負担の大きさによるものだと考えられる．
2. **ワードサラダの精度の影響** マルコフ連鎖による文章合成の結果によっては，たとえば「アメリカは完全に踏み切った。」のように自然に読めてしまう文書が偶然に出力される．今回の実験では不自然な文として提示する文書からもそれらを排除していない．こういった自然に読めてしまう不自然な文の存在により，被験者に混乱を招いたために精度に影響が生じたと考えられる．
3. **コーパスによる影響** 実験では，ニュース記事や Wikipedia の本文をコーパスとした．コーパスからの一部切り出しを自然な文として提示したが，その中には前後のつながりが不明で不自然に見えてしまう文もあった．また，コーパスのサイズを 10,000 文字程度と小さく限定したため，ワードサラダの精度にも影響を与えたと思われる．

不注意な被験者の影響については，提案方式のパフォーマンスの改善によってある程度軽減できる可能性がある．ワードサラダの精度の影響の改善は，より不自然な文のみを限定して合成できる方式の検討が必要である．コーパスによる影響については，コーパスサイズや文体等，提案方式に適切なコーパスを選ぶ必要がある．

[*4] タイ語には文末の記号が存在せず，さらに分かち書きされていない言語であり，形態素解析は困難な言語である．

表 8 機械校正の有無についての判別精度の比較 $P(Y = S|X = S)$
Table 8 Comparison of accuracy between with or without detection machine $P(Y = S|X = S)$.

	$n = 1$	標準偏差
$W = t$	0.85	0.156
$W = f$	0.87	0.125

表 9 被験者による平均正答率 $P(Y = S|X = S)$ と $P(Y = H|X = H)$ の差
Table 9 Comparison of accuracies in terms of number of subjects.

	$n = 2$	$n = 3$	自然な文
3 名の平均正答率	0.47	0.20	0.90
14 名の平均正答率	0.32	0.23	0.79

5.11 機械校正についての人間の正答率の影響

5.5 節では提案方式について，機械校正を用いた攻撃に対する安全性を評価した．しかし，機械校正が行われる場合とそうでない場合の文書について，人間が判別する際にどの程度の精度の差が生じるかは不明である．そこで，機械校正が行われた場合とそうでない場合との精度の違いを評価するために以下の実験を行う．

日本人学生 11 名に対し評価データを提示し，正答率を計測する．評価データは実験 5 と同じ方法で合成する．ただし，$n = 1$ のワードサラダについては Microsoft Word 2007 による文章校正が行われる文書のみを問題として提示する．

表 8 にこの実験の正答率と実験 5 の日本人の正答率の比較を示す．機械校正の有無について正答率の差は，0.02 であり，標準偏差の大きさを考慮すると誤差の範囲内である．すなわち，機械校正の人間に対する影響はないと結論付ける．

また，実験 5 では被験者が 3 名と少なく信頼性に欠ける懸念があった．今回の 11 名の結果を加えた 14 名の場合の平均の正答率を**表 9** に示す．被験者数が増えるに従って

表 10 提案方式で用いる言語を学んでいない攻撃者の判別率
Table 10 Detection rate made by attacker without knowledge of language of test sentence.

$n=1$	$n=2$	$n=3$	自然な文
0.41	0.37	0.29	0.75

正答率は変化しているが，階数 n に対する関係は変わっていないことが分かった．より大規模な被験者による実験は今後の課題に残すが，同様の結果が得られることが予想できる．

5.12 攻撃者の言語知識による影響

実験 4 では，効果の高いリレーアタックの精度を明らかにするための実験を行った．日本語の知識のある留学生について安全性が証明できれば，より知識の不足した一般の外国人に対しても安全であると帰着できる．しかし，まったく知識のない攻撃者による評価は行っていない．

そこで，提案方式で提示する文書の言語を学んでいない攻撃者の精度を測るための実験を行う．実験内容は，実験 5 で用いた中国語の評価データを日本人学生 10 名に提示し，その正答率を計測するものである．**表 10** に実験結果を示す．

実験結果では $n=1$ のときの $P(Y=S|X=S)$ は 0.41 となりランダムに回答した場合の正答率を下回った．対して自然な文についての正答率は 0.75 と比較的高い値となった．これについて被験者からは，「問題文が読めず，不自然さを感じられないために自然と回答した」との意見が得られた．

以上の結果から，この場合の攻撃者の 1 題あたりの成功率は 58%となることが期待される．5.6 節で安全性評価のために用いた留学生の 1 題あたりの CAPTCHA 成功率 P_{mc} は 75%である．よって，知識の不足した攻撃者による提案方式の成功率は実験 4 よりも低く，すなわち，より安全であることが明らかになった．

5.13 パフォーマンス改善についての検討

提案方式では 1 回の試行に 151.7 秒かかることが予想され，ユーザにかかる負荷がとても高く，現状の一般的な CAPTCHA と比べて運用上の問題がある．

提案方式はリレーアタックに対する耐性の高さを特徴としている．試行に多くの時間がかかることについて，リレーアタックの頻度を下げる効果や，正規ユーザとリレーアタック攻撃者の応答時間の差を用いた，時間の閾値の設定による精度の上昇が見込める等のメリットもあると考えられる．

また，実験では問題となる文書を 1 題ずつ順次提示しており，文書の文字数の制限も設けていなかったことがパフォーマンスに影響を及ぼした．そのため，文字数を読みやすい量に制限し，すべての問題を 1 度に提示した場合のテストを行ったところ，平均応答時間は 50 秒程度となり，大きく改善された．**図 8** にテストに用いた Web ページを示す．このようなインタフェースの工夫を行っていくことで，将来的に提案方式のパフォーマンスをより改善することができると考えられる．

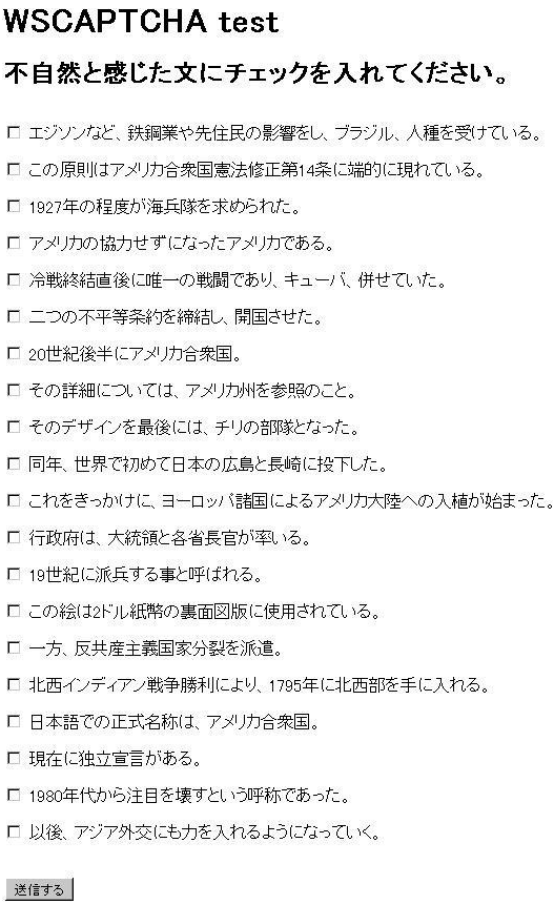

図 8 パフォーマンス向上のためのユーザインタフェース
Fig. 8 User interface to improve performance.

6. 結論

本稿では，合成された文章の不自然さを利用した CAPTCHA を提案し，3 つの攻撃モデルに対する安全性を評価した．また，提案手法の日本語以外の適用について検討し，適切な形態素解析がその条件となる見込みを得た．問題提示方法や文章合成手法の調整によるパフォーマンスの向上，文章校正以外の自然言語処理を用いた攻撃への耐性の検討，他言語への適用についての再評価を今後の課題とする．

謝辞 本稿の執筆に際してご助言いただいた，東海大学情報理工学部情報科学科内田理准教授に深く感謝いたします．また，様々なご指導をいただきました査読者に感謝いたします．

参考文献

[1] The Official CAPTCHA Site, available from ⟨http://www.captcha.net⟩.
[2] Yan, J. and Ahmad, A.S.E.: Breaking Visual CAPTCHAs with Naive Pattern Recognition Algorithms, *2007 Computer Security Applications Conference*, pp.279–291 (2007).
[3] Elson, J., Douceur, J., Howell, J. and Saul, J.: Asirra: A CAPTCHA that exploit interest-aligned manual image categorization, *ACM CSS 2007*, pp.366–374 (2007).
[4] Golle, P.: Machine Learning Attacks Against the ASIRRA CAPTCHA, *2008 ACM CSS*, pp.535–542 (2008).
[5] 山本 匠, Tygar, J.D., 西垣正勝：機械翻訳の違和感を用いたCAPTCHAの提案, 情報処理学会研究報告, CSEC-46 No.37 (2009).
[6] 山本 匠, Tygar, J.D., 西垣正勝：機械翻訳CAPTCHA（その2），コンピュータセキュリティシンポジウム2009論文集，pp.211–216 (2009).
[7] 鈴木徳一郎，山本 匠，西垣正勝：リレーアタックに耐性をもつCAPTCHAの提案，情報処理学会研究報告，CSEC-48, No.16 (2010).
[8] Geyer, J.: Practical Markov Chain Monte Carlo Charles, *SCI 1992*, Vol.7, No.4, pp.473–483 (1992).
[9] Issac, B.: Improved Bayesian Anti-Spam Filter Implementation and Analysis on Independent Spam Corpuses, *ICCET 2009*, Vol.2, pp.326–330 (2009).
[10] Zaragoza, H., Gallinari, P. and Rajman, M. (Eds.): Learning to Filter Spam E-Mail: A Comparison of a Naive Bayesian and a Memory-Based Approach, *PKDD 2000*, pp.1–13 (2000).
[11] Larvergne, T. et al.: Detecting Fake Content with Relatine Entropy Scoring, *CEVR*, Vol.377, pp.27–31 (2008).
[12] 森本浩介，片瀬弘晶，山名早人：N-gramと離散型共起表現を用いたワードサラダ型スパム検出手法の提案，情報処理学会研究報告，DBS-148, No.24, pp.1–8 (2009).
[13] MeCab: Yet Another Part-of-Speech and Morphological Analyzer, available from ⟨http://mecab.sourceforge.net⟩.
[14] ICTCLAS, available from ⟨http://www.ictclas.org⟩.

付　　録

A.1 実験5で用いたワードサラダの例

表 A·1 実験5．日本語，英語についての例題
Table A·1 Sample sentences in English and Japanese used in Experiment 5.

n	Japanese
1	朝鮮戦争が積極的に起こる第二次世界大戦が 完全にはニューヨークに発効されて構成され、 自由のコントラ支援した。
3	大戦以前は非戦争時にはGDPに対する軍事費の 比率が低い国だったが、大量破壊兵器は見つからず 石油を狙った侵略行為と批判する声があがった。
Hum	後にアメリカ人は「明白な天命」をスローガンに 奥地への開拓を進め、たとえ貧民でも自らの労働で 土地を得て豊かな暮らしを手に出来るという文化を 形成して「自由と民主主義」理念の源流を形作っていった。
n	English
1	With the vast bulk of Englishmen and the Louisiana territory separated from the Southwest, which states were organized on September 17, the 100th century, has been described.
3	In 1507, German cartographer Martin Waldseemuller produced a world map on which he named lands of the Western Hemisphere America after Italian explorer and cartographer Amerigo Vespucci.
Hum	The United States of Americasis a federal constitutional republic comprising fifty states and a federal district.

表 A·2 実験5．中国語についての例題
Table A·2 Sample sentence in Chinese used in Experiment 5.

中国語↓ S(n=1)↓ 这些数据均与人类在该州获得参议院三分之二通过后，在长岛等地挑选和山谷，↓ 最后终于废除了朝鲜战争后，并且对当地动植物保护区域全部加起来高达14,000英尺。↓ ↓ S(n=3)↓ 经历独立战争后，美国获取加利福尼亚州、内华达州、↓ 犹他州全部地区，科罗拉多州、亚利桑那州、新墨西哥州和怀俄明州部分地区。↓ ↓ H↓ 1607年，位于伦敦的弗吉尼亚公司在北美切萨皮↓ 克湾的詹姆斯敦建立英国的第一个短暂殖民地。←

推薦文

ワードサラダと呼ばれる，コンピュータによって自動生成された「文法は正しいが意味が支離滅裂である文章」と，そうではない文章を人間に判断させることでCAPTCHAを実現する手法が従来提案されている．本稿ではこの手法に関して，Microsoft Word等の校正ツールを通すことで，不自然であるかどうかを判定可能であると仮定し，その際の安全性について厳密な評価を行っている．安全性の評価や対策手法，多言語に対する適用等，よく検討されている．

評価の信頼性は高くさらなる発展も望めることから，論文化の価値は高いと考える．

（コンピュータセキュリティ研究会主査　松浦幹太）

鴨志田 芳典　（正会員）

2013年東海大学大学院工学研究科情報理工学専攻修了．2013年(株) TOKAIコミュニケーションズに入社．システムイノベーションサービス本部所属．在学中，情報セキュリティ，ウェブサービスに関する研究に従事．2012年マルチメディア，分散，協調とモバイルシンポジウム（DICOMO2012）優秀論文賞受賞．

菊池 浩明　（フェロー）

1988年明治大学工学部電子通信工学科卒業．1990年同大学院博士前期課程修了．1994年同博士（工学）．1990年（株）富士通研究所入社．1994年東海大学工学部電気工学科助手．1995年同専任講師．1999年同助教授，2000年同電子情報学部情報メディア学科助教授，2006年同情報理工学部情報メディア学科教授．2008年同情報通信学部通信ネットワーク工学科教授．1997年カーネギーメロン大学計算機科学学部客員研究員．2013年明治大学総合数理学部先端メディアサイエンス学科教授．WIDEプロジェクト暗号メールシステムFJPEMの開発，認証実用化実験協議会（ICAT），IPA独創情報技術育成事業等に従事．暗号プロトコル，ネットワークセキュリティ，ファジィ論理，プライバシ保護データマイニング等に興味を持つ．1990年日本ファジィ学会奨励賞，1993年情報処理学会奨励賞，1996年SCIS論文賞，2010年情報処理学会JIP Outstanding Paper Award．2013年IEEE AINA Best Paper Award．電子情報通信学会，日本知能情報ファジィ学会，IEEE，ACM各会員．