文章合成の不自然さの評価と応用

Application and Evaluation of Synthesized Sentences

[1] 鴨志田　芳典,　　[1] 菊池　浩明
[1]Yoshifumi Kamoshida,　[1]Hiroaki Kikuchi
[1] 東海大学
[1]Tokai University

Abstract: Artificially synthesized sentences are used for malicious purposes such as unsolicited commercial junk submission to Web site. We study a problem to distinguish between natural and synthesized messages generated in Markov chains and show experimental results. Based on the difficulties of the problem, we consider a new application to CAPTCHA, a type of challenge-response test used in computing to ensure that the response is not generated by a computer.

Keywords：words alada, Markov Chain, captcha, Synthesied 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伴う不正行為への対策として CAPTCHA(Completely Automated Public Turing Test To Tell Computers and Humans Apart) と呼ばれる機械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事でボットやエージェントなどのプログラムされた入力と人による入力とを識別する．しかしながら通常広く用いられている図 1 の例のような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事が J. Yan らによって報告されている [2].

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事の困難さを利用した CAPTCHA で，画面上に表示された複数の画像から犬か猫かを人間に選択させる．マルウェアによる解析は困難とされていたが，P.Golle により画像の特徴から犬と猫を判別する事により問題を解決する手法が提案されている [4]．また山本らは，

図 1: Yahoo で使用されている CAPTCHA

つまり自分が、怒りに引き揚げても、謂わばいいくらいでしたのぞ》を食べなければ通俗の苦しみ、それは、子供のは爽快《もっ》のこぶしを感じるの腰布（しかし、めしを、もじもじした。

図 2: ワードサラダ合成例

コンピュータが文章の不自然さを理解する事の困難さを利用し，機械で繰り返し翻訳された文と翻訳前の文を判別させる CAPTCHA を提案している [5]．このように，CAPTCHA はいずれ解析される可能性があるため，方式の多様性が求められている．

本研究では言語を用いた CAPTCHA の利用性を評価するための最初の検討として，ワードサラダと呼ばれるマルコフ連鎖よる文章の自動合成を用いた 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章を合成する．太宰治 著「人間失格」から合成したワードサラダの例を図 2 に示す．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，人が見れば話題の繋がり方などから不自然とわかる．合成された文章は，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り，CAPTCHA として有効に利用できると考えられる．ワードサラダを CAPTCHA に応用する事の利点として以下の 3 点が挙げられる．

(i) 機械的に多量の問題文生成が可能である．
(ii) 不自然さの程度を操作できる．
(iii) 外国の不正ユーザからの攻撃に耐性がある．

本稿では文章を用いた CAPTCHA の有効性を実験データから評価する．まず，提案方式に用いる文章合成のアルゴリズムを解説し，それを CAPTCHA に利用する手法を提案する．その後ワードサラダの不自然さを評価するために行った研究 [10] の実験データから提案手法の有効性の検証を行い，その結果を報告する．

2 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 の確率は以下の条件付き確率に従う．

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

日本語は分かち書きされていないため，日本語の単語単位の N-gram 頻度を得る為には前処理として形態素解析器による分かち書きが必要である．抽出した単語 N-gram 頻度データ基に 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．前述の様に 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事が可能である．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．

3 提案手法

3.1 概要

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一つずつ提示し，設定されたしきい値以上の精度で正しく答えられるかどうかで人と機械を判別する．

3.2 方法

提案手法はコーパスの収集，マルコフ連鎖モデルの作成，問題となる文の合成，CAPTCHA による認証という構成からなる．それぞれを以下で述べる．

3.2.1 コーパスの収集

以下の 3 つのコーパスを用いる

1 新聞社の最新ニュース記事

2 新聞のアーカイブ等の有料コンテンツ

3 SNS の日記等の口語体の文章

文章合成の為のコーパスは 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する問題では，ボットはそれを Web で検索することで判断可能である．よって，関連研究 [6] の様に新聞のアーカイブ等の有料コンテンツや青空文庫等，Web 検索にかからない文章をコーパスとする．また，口語表現や文法の間違った文章をコーパスとした場合形態素解析の精度が落ちる為，より不自然なワードサラダが合成できる．Web に公開されていない学生の書いたレポートの「ですます」調と「である」調の混在した文章や，mixi 等の非公開型の SNS の日記などを用いる事も検討する．

3.2.2 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 $N=1$ でのワードサラダ合成を行う．文章の不自然さと言う点では単語をランダムに結んだ物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事が予想できる．ワードサラダは品詞の並びが文法として適切となる特徴がある．

3.2.3 問題となる文の合成

CAPTCHA による認証を行う為に提示する問題として使用する自然な文とワードサラダをそれぞれ $c_1, c_2, \ldots, c_n$ と $a_1, a_2, \ldots, a_m$ とし，$n+m$ を s と定義する．n，m はそれぞれ，問題として使用する自然な文の数とワードサラダの数であり，s は 1 度の CAPTCHA で提示する問題の総数である．ワードサラダ a_i は 2 章で解説したマルコフ連鎖による文章合成を用いて合成する．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．c_i は収集した Web 検索にかからないコーパスの一部の文を利用する．

3.2.4 CAPTCHA による認証

合成した文を一題ずつランダムに提示し，ユーザは提示された文に対し「自然」か「不自然」を選択する．自然な文とワードサラダとを正しく解答した回数を正解数 k とする．k が閾値 θ 以上ならば CAPTCHA 成功とする．

3.3 機械による攻撃への耐性

3.3.1 総当り攻撃

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，全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．攻撃者が n と m の割合を知らないと仮定した場合，問題は自然か不自然かの 2 択であるため，1 つの問いに対して正解する確率は 1/2 であり，総当り攻撃で k 問正解する確率は試行回数 s, 成功数 k, 確率 1/2 の二項分布で求められる．よって，総当り攻撃が $k \geqq \theta$ を満たす確率 P_r は,

$$P_r = \frac{1}{2^k} \sum_{k=\theta}^{s} \binom{s}{k}$$

となる．これを FAR(False Machine Acceptance Rato) と定義する．例えば $n + m = 15$, $\theta = 13$ の時，総当り攻撃が成功する確率は約 0.37% である．

3.3.2 人による攻撃 (リレーアタック)

近年 CAPTCHA に対しての攻撃において，高機能なマルウェアによる攻撃の他に，人間を使ったリレーアタックと呼ばれる攻撃が問題になっている [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 を解く人間は発展途上国の低賃金労働者である．提案手法は文の不自然さを判断できる程度の言語能力がユーザに問われるため，CAPTCHA を行う正規ユーザの母国語を問題とした場合を想定した閾値 θ を設定することにより，正規ユーザの母国以外の不正ユーザからのリレーアタックをある程度防ぐ事が可能であると考えられる．

4 評価

4.1 実験

人がワードサラダの不自然さの判別を行った際の精度と応答時間を評価するために行った研究 [10] の実験データから，提案手法の有効性を評価する．評価データは収集したコーパスから合成した不自然な文章とする階数 $N = 1, \ldots, 3$ のワードサラダと，自然な文章とする合成に用いた文章の一部である．実験 1，実験 3 では評価データは実験に用いたワードサラダは 5000 文字程度の政治・経済に関する記事から合成し，実験 2 では青空文庫から収集した 4 つの文章から抜き出した 20000 文字程度の文章から合成した．

実験 1．日本人による主観実験 情報系の学生 8 名に対し次の主観実験を行った．5 行からなる評価データを $n = 50$,$m = 50$ の計 100 題を提示し，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，その正答率と応答時間を計測した．

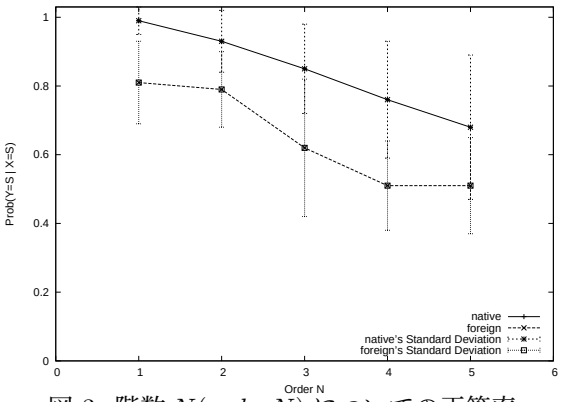

図 3: 階数 $N(orderN)$ についての正答率

実験 2．留学生による主観実験 日本語を学んだ留学生と日本人学生との間にどれ程の日本語の識別能力の違いがあるかを検証するため，日本語を学んだ留学生 3 名に対し，実験 1 と同様の実験を行った．ただし，自然な文章に対しての評価は行っていない．

実験 3．提示する文章の量による実験 一度に提示する文章の量による精度への影響を確認するための実験を行う．7 名の被験者に対し，予め 100 題ずつ用意された 1 行の評価データを自然な $n = 5$,$m = 10$ の割合でランダムに 15 回提示する．ワードサラダの階数 $N = 1, \ldots, 3$ について，それぞれの場合ごとにその文が自然か不自然かを判断させその正答率と応答時間を計測する．実験は一人につき複数回行い，13 件のデータが得られた．

実験 4．従来手法での主観実験 従来の文字列画像を変形させた CAPTCHA と提案手法の精度の比較を行うため，[1] に提示されている CAPTCHA のデモプラグラムを利用し，2 名に 20 回ずつ CAPTCHA を行ってもらい，CAPTCHA 成功率と必要時間を計測した．

4.2 実験結果

実験 1，2 でのワードサラダに対する正答率と応答時間の平均をそれぞれ図 3 と図 4 にそれぞれ示す．日本人の自然な文に対する正答率は 59%であり，応答時間は 21.07 秒であった．以降，留学生の自然な文に対する正答率を 50%と仮定する．

実験 1 から得られた結果に対し， X を入力を表す確率変数, Y を出力を表す確率変数, H を人間による文章,

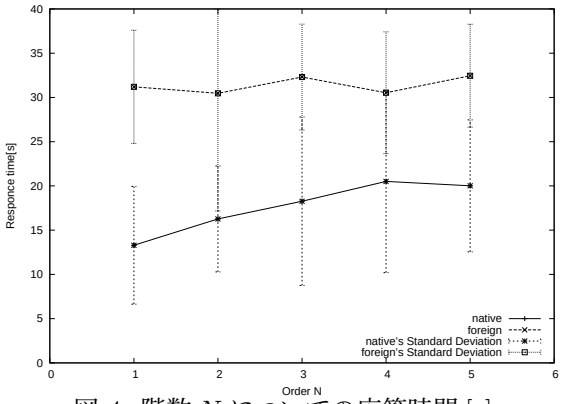

図 4: 階数 N についての応答時間 $[s]$

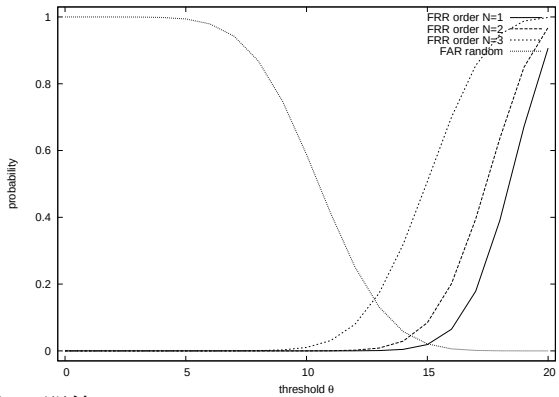

図 5: 閾値 θ についての FRR と $FAR(n = 5, m = 15)$

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確率は $P(Y = H|X = H)$ と表せる．実験 1 から得られた $N = 1$,$n = m$ 時の解答の正答率は，条件付確率 $P(Y|X)$ として表 1 の様に与えられた．

表 1: $N = 1$, $n = m$ の時の条件付確率 $P = (Y|X)$

入力文書 \ 判別文書	$Y = H$	$Y = S$
$X = H$	0.59	0.41
$X = S$	0.01	0.99

自然な文章とスパム文章を出題する確率 (事前確率) はそれぞれ,

$$\begin{aligned} P(X = H) &= \frac{n}{s} \\ P(X = Y) &= \frac{m}{s} = 1 - \frac{n}{s} \end{aligned}$$

である．従って，自然な文章とスパム文章の歪みを考慮した CAPTCHA 成功率は，これらの同時確率 P(X, Y) で次のように与えられる．

$$\begin{aligned} P(Y = H, X = H) &= P(Y = H|X = H)P(X = H) \\ P(Y = S, X = H) &= P(Y = S|X = H)P(X = H) \\ P(Y = H, X = S) &= P(Y = H|X = S)P(X = S) \\ P(Y = S, X = S) &= P(Y = S|X = S)P(X = S) \end{aligned}$$

CAPTCHA の検査に失敗するには，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，CAPTCHA 失敗率 P_q を以下のように定める．

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

このとき，s 回の CAPTCHA の検査に k 回誤答する確率は，確率 P_q の 2 項分布で表すことができる．

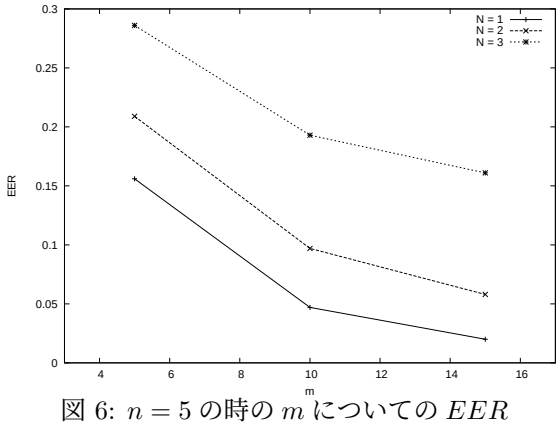

図 6: $n = 5$ の時の m についての EER

3.3.1 節で求めた総当り攻撃による CAPTCHA 試行が $k \geqq \theta$ となる機械受入れ率 FAR に対し，人間が CAPTCHA を試行したのに $k < \theta$ となる確率をとなる確率を人間拒否率 FRR(False human Rejection Rate) と定義する．FRR と FAR は次の式で与えられる．

$$\begin{aligned} FRR &= \sum_{k=\theta}^{s} \binom{s}{k} {P_q}^k (1 - P_q)^{s-k} \\ FAR &= \frac{1}{2^k} \sum_{k=\theta}^{s} \binom{s}{k} \end{aligned}$$

また，$FAR = FRR$ となる値を EER(Equal Error Rate) とする．

$n = 5$，$m = 15$ の場合に，階数 $N = 1, \ldots, 3$ のワードサラダにおいての閾値 θ について，FRR と FAR を図 5 に示す．$n = 5$ の時の $m = 5, 10, 15$ のそれぞれの場合において，階数 N についての EER を図 6 に示す．EER の最も低くなる $N = 1$ のワードサラダにおいて，ワードサラダを増やすことによる CAPTCHA 精度の変化を図 7 に示す．また (階数 $N = 1, n = 5, m = 15$) の時の留学生との CAPTCHA 精度の図 8 に示す．

実験 3 での正答率を表 2 に，応答時間を表 3 にそれ

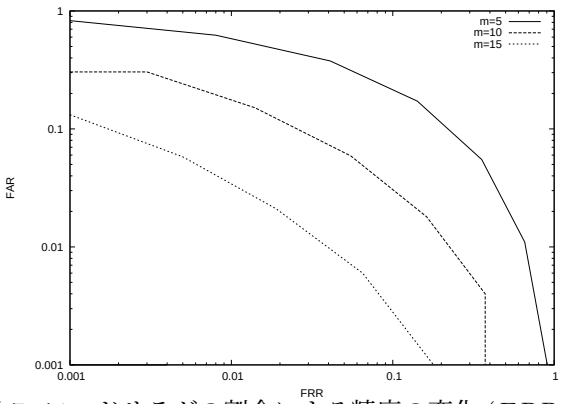

図 7: ワードサラダの割合による精度の変化 (FRR についての FAR)

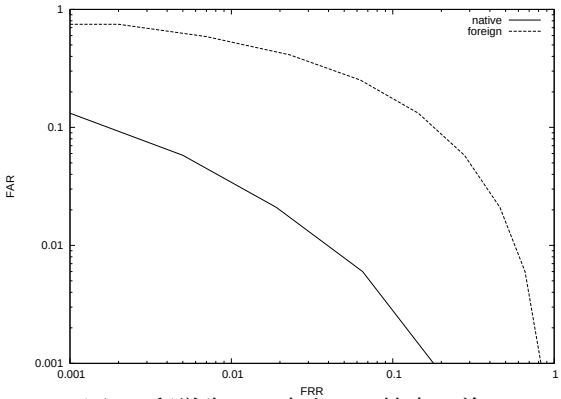

図 8: 留学生と日本人との精度の差

表 2: 実験 3:文章量が 1 行の時の正答率

	階数 $N=1$	$N=2$	$N=3$
$P(Y=H\|X=H)$	0.91	0.80	0.68
$P(Y=S\|X=S)$	0.73	0.62	0.45

表 3: 実験 3:文章量が 1 行の時の応答時間 $[s]$

種類	階数 $N=1$	$N=2$	$N=3$
自然な文	$8.05[s]$	$8.12[s]$	$7.44[s]$
ワードサラダ	$6.19[s]$	$7.76[s]$	$8.58[s]$

ぞれ示す.

実験 4 での CAPTCHA 成功率は 100%で，平均応答時間は $9.74[s]$ であった.

4.3 考察

4.3.1 実験結果からの考察

実験 1，2 との比較から，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事がわかる．また，階数 N に依存して正答率が減少する事からワードサラダは階数を増やす程自然な文と判別が付きづらくなると言える．図 5 より，階数 $N=1$ のとき，もっとも EER に近くなる θ の値は $\theta=15$ である．すなわち，$s=20$ 個の文章の 3/4 をワードサラダとすることで，本人拒否率と機械受入れ率を最小化できることを意味している．

図 6 と図 7 より，問題の中に不自然な文章を多く混ぜる程 CAPTCHA の精度が上がる事がわかる．図 8 より，閾値 θ を日本人の EER となる $\theta=15$ と設定した時，提案手法は正規ユーザの母国以外の不正ユーザからのリレーアタックをある程度防ぐ事が可能であると言える．

実験 3 の結果では，文章量を減らす事により全ての場合において自然な文を判別できる確率が増加した．人間は文章量が多いと文を不自然と判断する傾向にある．また表 2 より，同時に提示するワードサラダの階数 N の増加に伴い，自然な文を判別できる確率は減少している．自然な文は変えていない事から，同時に出題する不自然な文に十分な不自然さを確保できれば，自然な文を判別できる確率も上がるものと思われる．

以上の実験結果より，CAPTCHA として適切なパラメータは，$n=5$，$m=15$ の 20 題の問題で行い, 正解数 k の閾値 $\theta=15$ であり，その時の FRR 及び FAR は 2%となる事が期待できる．

しかし，実験 4 の結果より従来手法の平均応答時間は 9.74[s] であるのに対し，図 4 より階数 N=1 のワードサラダの平均応答時間は 13.3 秒，自然な文の平均応答時間は 21.07 秒である．自然な文を 5 題，ワードサラダを 15 題出しているため，CAPTCHA にかかる合計時間はおよそ 307.85 秒と予想される．従来の文字列画像を変形させた CAPTCHA のおよそ 30 倍の時間がかかり，ユーザにかかる負荷はとても高い．

4.3.2 提案手法の問題点と改善案の検討

今回の実験では，関連研究 [10] で行った実験結果から CAPTCHA に応用する際の再評価を行った．実験で使った問題は自然な文章を合成しやすくなるようにコーパスの規模を 5000 文字から 20000 文字と非常に少ないものにしている．そのため自然なワードサラダも多く出力され，FAR は高くなっていると思われる．また，充分な精度を得るために文章を 5 行ずつ提示しており，ユーザに掛かる負荷は高い.

ワードサラダ合成に用いるコーパスをより大きくする事で，より不自然な文章が合成できることを期待できる．また，出現頻度の低い単語へのマルコフ連鎖の遷移確率を高くする事により，不自然な文章を合成で

きる確率が高まるものと思われる．しかしその場合，文として成立するかどうかは検討の余地がある．ワードサラダに充分な不自然さが与える事が出来れば，文章量を減らし，パフォーマンスを上げる事も可能である．

ワードサラダに対する検出手法を応用する事で，FAR を下げる事ができる．関連研究 [8] は Google-N-gram 頻度データを利用し，カルバック・ライブラー情報量によりワードサラダのスコアリングを行い，ワードサラダの検出を行う手法である．また関連研究 [9] は [8] の手法に加え wikipedia 本文のスナップショットから抽出した離散共起表現を利用し，ワードサラダの検出を行う．どちらの手法も，ワードサラダを高い精度で検出する事が確認されている．これらの手法を提案方式で作成したワードサラダに応用することにより，一定の不自然さが保証されたワードサラダを自動的に合成する事が検討できる．

これらのワードサラダの検出手法を用いて提案方式へ攻撃が行われた場合，FRR は非常に低くなると考えられる．しかし，どちらの手法もワードサラダのスコアリングには大規模のコーパスと大量の計算量を必要とするため，システムの処理時間が膨大になりパフォーマンスは低くなると思われる．

5 おわりに

本論文では，マルコフ連鎖により合成された文章の不自然さを応用した CAPTCHA の提案を行い，ワードサラダの不自然さを評価する実験データからその性能を評価した．20 題の文章中に階数 N=1 のワードサラダ 15 題と自然な文 5 題という最も精度の良くなる条件下では，人間拒否率及び機械受入れ率 2%の精度と，必要時間 308.75 秒のパフォーマンスで認証を行う事が可能である事を明らかにした．より短い文章量で不自然さを保証したワードサラダ合成方法の検討と，提案方式の実装と実験を今後の課題とする．

参考文献

[1] The Official CAPTCHA Site, (http://www.captcha.net)

[2] J. Yan and A. S. E. Ahmad: Breaking Visual CAPTCHAs with Naive Pattern Recognition Algorithms, 2007 Computer Security Applications Conference, pp. 279-291，2007.

[3] J. Elson，J. Douceur，J. Howell and J. Saul, Asirra: a CAPTCHA that exploit interest-aligned manual image categorization, 2007 ACM CSS, pp. 366-374, 2007.

[4] P. Golle: Machine Learning Attacks Against the ASIRRA CAPTCHA, 2008 ACM CSS, pp. 535-542 2008.

[5] 山本匠，J. D. Tygar, 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，J. D. Tygar，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 (2009.10)

[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] T. Larvergne, et al.,：Detecting Fack Content with Relatine Entropy Scoring', CEVR, Vol.377, pp. 27-31, 2008.

[9] 森本，片瀬，山名：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.

[10] 鴨志田芳典，菊池浩明：マルコフチェインによるワードスパムの合成実験とその評価について，第 72 回情報処理学会全国大会，講演番号 2G-1，2010.

[11] MeCab，MeCab: Yet Another Part-of-Speech and Morphological Analyzer, (http://mecab.sourceforge.net/)

連絡先
東海大学菊池研究室
E-mail: syake@cs.dm.u-tokai.ac.jp