C8　文章合成の不自然さを用いたCAPTCHA

発表者 0BDRM013 鴨志田 芳典
指導教員 菊池 浩明 教授

The study of CAPTCHA using artificial synthesis sentences.

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.

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通常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や大量の人手を用いたリレーアタックによって破られてしまう [1].

本研究では，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．ワードサラダにより合成された文章は、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．

2　提案手法

提案手法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一つずつ提示し，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．h，s をそれぞれ，自然な文の数とワードサラダの数とし，$h+s$ を c と定義する．自然な文は収集したコーパスから一部の文を利用する．

認証プロセスでは，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 とし，k が閾値 θ 以上ならばCAPTCHA成功とする．業績1で行った実験では，提案手法はリレーアタックや総当たり攻撃に対して耐性を持つ事を示した．しかし，CAPTCHAに掛かる時間は308秒となり，パフォーマンスは低い事がわかっている．

3　評価

3.1　実験1

提示する文章量による精度とパフォーマンスの変化を評価するために，以下の実験を行った．情報系の学生7名に対し，1文からなる評価データを $h=5$, $s=15$ の計15題を提示し，実験1と同様の方法で正答率と応答時間を計測した．

3.2　実験結果

実験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，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 $n=1$ の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．また，平均応答時間は7.43秒であった．

Table 1: 実験2:$N=1$ の時の条件付確率 $P=(Y|X)$

入力文書 \ 判別文書	$Y=H$	$Y=S$
$X=H$	0.91	0.09
$X=S$	0.27	0.73

3.2.1　解析1

自然な文章とスパム文章を出題する確率はそれぞれ，$P(X=H)=\frac{h}{c}, P(X=Y)=1-\frac{h}{c}$ である．

c 回のCAPTCHAの検査に k 回誤答する確率は，CAPTCHA失敗率 P_q の2項分布で表すことができる．人間がCAPTCHAを試行したのに $k<\theta$ となる確率を，人間拒否率 $FRR=\sum_{k=\theta}^{c}\binom{c}{k}{P_q}^k(1-P_q)^{c-k}$ として与える．同様に攻撃者と仮定する機械による総当り攻撃が $k \geqq \theta$ を満たす確率を FAR と与え，$FAR=FRR$ となる値を EER とする．

過去研究との結果の比較を図1に示す．図1より，文章量について精度の差はほぼ見られない事ことから，提案方式で提示する文章量は1文が望ましいと言える．CAPTCHAに必要な時間はおよそ152秒と予想された．

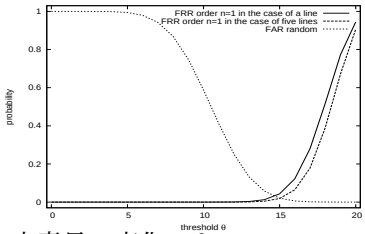

Fig. 1: 文章量の変化による θ についての精度

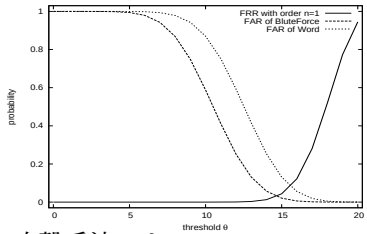

Fig. 2: 攻撃手法による FRR についての FAR

3.2.2 解析 2

文章校正による検出を用いた攻撃を行われた場合について検討する．提案方式について，一つの文について文章校正が行われる確率を $P(w)$ とする．実験 1，実験 2 で使用した評価データについて MicroSoft Word による文章校正を 1 題ずつ行い，条件付確率 $P(w|X = S) = 0.24$，$P(w|X = H) = 0$ を得た．

基本解析より得られた最適な条件で提案手法を行った場合，$P(X = H) = 0.75$，$P(X = S) = 0.25$ となり，$P(w) = 0.06$ で検出が行われる．この時，検出が行われた文の入力がスパムである確率は，条件付き確率で $P(X = S|w)$ と表せる．

これを元に，表 2 に示す文章校正を用いた検出によるスパム判定機を得る．この判定機による機械による攻撃の成功率は，条件付確率 $P(Y|X)$ として表 3 の様に求められた．

Table 2: 文章校正を用いた検出によるスパム判定機

入力文書 \ 判定	$\overline{w}$	w
$X = H$	0.6	0.0
$X = S$	0.4	1.0

Table 3: 判定機を用いた機械の $P = (Y|X)$

入力文書 \ 判別文書	$Y = H$	$Y = S$
$X = H$	0.564	0.436
$X = S$	0.436	0.564

この攻撃の成功率を P_m とし，この攻撃による機会受けれ率を $FAR_w = \sum_{k=\theta}^{s} \binom{s}{k} {P_m}^{s-k}(1 - P_m)^k$ とする．正解数 k の閾値 θ についての FAR_w と FRR を図 2 に示す．基本解析と同様の条件で提案方式を行い，機械により文章校正ツールを用いた検出が行われた時，EER は 8%となり，精度は半減する．

4 提案方式の他言語への適用

4.1 実験 2

提案方式が他言語でも適用可能か検証する為に，以下の実験を行った．日本人学生 3 名，英語，中国語，タイ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，正答率を計測した．評価データは，Wikipedia のアメリカ合衆国の記事の本文から抽出し合成した．

4.2 実験結果

実験 2 の結果を表 2 に示す．結果から，各言語とも階数 n の値が低い時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った．タイ語についてはの文末の学習が上手く出来なかった為，全ての場合で不自然であるという結果になった．

Table 4: 実験 3:各言語毎の不自然な文の判別精度

Language	$n = 1$	$n = 2$	$n = 3$	Natural
Japanese	0.87	0.47	0.20	0.90
English	1.0	0.8	0.6	0.7
Chinese	1.0	0.8	0.5	0.7
Thai	1.0	1.0	0.8	0.6

5 おわりに

本稿では，合成された文章の不自然さを利用した CAPTCHA の提案を行い，その性能を評価した．

参考文献

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

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

業績リスト

1. 鴨志田, 菊池, ”文章合成の不自然さの評価と応用”, ファジィシステムシンポジウム 2010, pp.1069-1074, 2010.
2. 佐藤，長谷川，鴨志田，菊池，”印象に残りやすい日本語パスワード合成法について”，情報処理学会第 73 回全国大会, pp. 527-528, 2011.

他 1 件