K-049

インターネット上に湧出する文章の特徴と そのチューリングテストのバリアフリー化への利用

Grammatical and phonological features of sentences on Twitter in respect to utilization for CAPTCHA-like test for people with visual disability

山口 通智†‡　中田 亨‡　岡本 健†
Michitomo Yamaguchi　Toru Nakata　Takeshi Okamoto

1．まえがき

ネット上での個人認証手続きに利用される技術に、人間と人工知能による自動プログラム（ロボット）を判別するチューリングテストがある。このテストは、ロボットが大量のアカウントを不正に取得するような攻撃に対して効果的である。

テストには、人間には容易に解けるが、現在のロボットでは解くことが難しいと仮定できる AI(Artificial Intelligence、人工知能) 問題を利用する。このようなテストの代表例は、歪んだ文字画像を読み取らせる CAPTCHA (Completely Automated Public Turing Test To Tell Computers and Humans Apart) がある。画像の解釈を AI 問題としているため、視覚障碍者には利用しにくいことが指摘されている。

代替案として、変形した音声を利用する音声 CAPTCHA があるが、これは人間にも非常に難しいと報告 [1] されている。Google の音声 reCAPTCHA を例に挙げれば、現在は 20 秒もの音声に対する回答を要求している。しかしながら、これを解くのは人間にも事実上不可能であることが、Google Groups の reCAPTCHA フォーラムにて、開発者や利用者らに指摘されている。このことから、視覚障害者が現状の CAPTCHA を利用するには、高い障壁が存在する。

2．提案方式

2.1　テストの要件

本研究の目的は、知覚障碍によらず誰でも利用できる CAPTCHA 様のチューリングテストを考案することである。はじめに、その要件を列挙する。

(1)　「バリアフリー要件」：特定の知覚のみの使用に限定されないこと。
(2)　「識別性要件」：人間には容易に解けるが、現状のロボットには解答が難しい問題を生成できること。
(3)　「知識非依存性要件」：テストの難易度が特定の知識の有無に強く依存しないこと。
(4)　「問題新規性要件」：未使用で新規な問題を無数かつ自動で作れること。

2.2　文意文脈解釈問題

要件(1)、(2)を満たすため、本研究では文意文脈解釈問題を用いる。この種の問題は多種存在する [2] が、本稿では、自然な文とワードサラダ文を識別する問題を取り上げる。ワードサラダとは、「てにをは」といった文法構造は正しいが、登場する単語はランダムに選ばれているため、内容が不自然な文章である。ワードサラダは文法的には正しい文であるため、内容の自然さの識別には常識が必要になり、ロボットには難しいと期待できる。常識による判別であるため、要件(3)も満たすと考えられる。

以下に、テストの基本構成を示す。

(1)　「作問要求の受付」：利用者が認証などの手続きで、テストが課せられるべき状況に到達すると、作問プログラムが起動する。
(2)　「素材文章の収集」：問題文の素材になる文書の集合を取得する。
(3)　「作問」：素材文章から、自然な文とワードサラダ文を複数作成する。
(4)　「出題」：利用者に作成した問題文を提示し、特定の特徴を持った文を選ぶように指示する。問題は文字情報として提示できれば十分なので、ディスプレイやスピーカなど、利用者の障碍に合わせた出力機器が利用できる。
(5)　「回答と確認」：利用者は出題に対して回答する。作問プログラムは、回答が所期するものと一致すれば、テストに合格したと判定する。

2.3　ワードサラダの生成方法

ワードサラダの生成には、式**(1)**に示される N 階マルコフ連鎖モデルがよく利用される。この方式は、文中のある形態素は、直前 N 個の形態素により決定される連鎖型共起表現であるという仮定に基づく。具体的には、素材文章を形態素解析したものをコーパスとし、式**(1)**に従い組み合わせることで、コンピュータにより効率的に生成できる。

$$\Pr[\mathrm{X}_{n+1}=x|\mathrm{X}_n=x_n,\ldots,\mathrm{X}_0=x_0]= \Pr[\mathrm{X}_{n+1}=x|\mathrm{X}_n=x_n,\ldots,\mathrm{X}_{n-N+1}=x_{n-N+1}] \quad (1)$$

生成される文章は、N が小さいほど支離滅裂になり人間には解きやすい。例として、FIT2013 のウェブページに掲載されている「Welcome to とっとり」の紹介文を元に、N=1,6 のマルコフ連鎖モデルに従い生成したワードサラダを示す。N=6 の場合、結果的に素材文章の一部が切り出された形になり、自然な文になっている。これでは要件(2)を満たさない。

- N=1
  - 好天が住んで温暖で、これをはじめと、
  - 大山をはじめと、冬には、当時、春から秋は
- N=6
  - 秋は好天が多く、冬には降雪もあるなど、
  - 西端に位置する中国地方の北東部に位置し、

本研究では、離散型共起表現への対応という問題を指摘する。例えば、文中で「もし」の後には「なら（ば）」の出現頻度が高いと考えられる。しかし、2 つの語の間隔が長く空くこともある。長い間隔にも対処するには、N を大

† 筑波技術大学技術科学研究科
‡ 産業技術総合研究所セキュアシステム研究部門, RISEC

きくする必要があるが、ランダム性が弱くなる。離散型共起表現を用いたワードサラダの検出方式 [3] も提案されているので、このトレードオフは捨て置けない。

本研究では、形態素解析と構造解析を利用し、離散型共起表現を含めたコーパスを構成し、それを用いたワードサラダの生成を提案する。この利点を、文字列「もし／明日／の／午後／晴れ／た／なら／ば」を例に挙げて説明する。ここで '／' は、形態素の区切りを示す。構造解析を実行すると、連鎖型共起表現「明日の → 午後 → 晴れたならば」と離散型共起表現「もし → 晴れたならば」を得る。この例では、各形態素と構造解析結果の「もし晴れたならば」をワードサラダ生成用のコーパス として登録する。コーパスに登録した離散型共起表現を、あたかもひとつの形態素のごとく扱うことで、N = 1 のような小さい値のままで、離散型共起表現を含むワードサラダが生成できる。一方この例で、N 階マルコフ連鎖モデルのみで離散型共起表現を組み込むには、「もし」から「なら」までの間隔 (N = 6) への対応が必要になることから、提案方式の優位性は明らかである。

2.4 素材文章の選択と文章の改変

要件(3)を満たすため、素材文章は特定ジャンルに偏った文章ではいけない。また、要件(4)を満たすため、得られる文章量が十分に多い必要がある。よって本研究では、ネット上の文章を利用する。

ネット上の文章は公開情報である。素材文章をそのまま自然な文として使用すると、問題文と一致する文字列が検索でヒットするかどうかで、問題の回答をすることができてしまう。本研究では、この問題点を解決するため、文章の改変を 2.2 節の構成(3)と(4)の間におこなう。改変の方式は、漢字をカナに開いた後に、文字の追加／削除と置換を組み合わせておこなう。置換については、子音の交代をおこなう。これは方言などに見られる単語の子音の違いを指す。改変率が一定以上ならば、改変後の文章から改変前のものを探すのは困難であると考えられる。改変前後の形態素解析は一致しないため、「もしかして検索」のような形態素ごとの検索語修正に対しても効果的である。

最後に、四者択一の回答方式で、ワードサラダ文を選択させる問題を、作問例として示す。

(ア) ナナジスジニユウヒョクヲタベタ。
(イ) モヒハレタナラユウエロチニコマッタ！
(ウ) オガネガタリズホロガガエナイ。
(エ) ヤクタイノナイコトヲガロガエテイルト、
正解は（イ）である。

3．素材文章に Twitter を用いた場合の検討

3.1 Twitter 文の特徴

Twitter では新規文章が頻繁に生成されるので、要件(4)によく適合する。本稿では、2013 年 5 月 15 日 19 時 からの約 8 時間に、API を介して収集した識別可能な 13713 件のツイートについて解析する。文の漢字をカナに開き、顔文字などの特殊記号を削除したのちに文字数を数えた結果得られた分布の平均値は 32.2 文字であり、$\lambda = 0.03$ の指数分布でよく近似できた。指数分布で近似できるので、ツイート相互の文字数の相関は一般に弱く、おおむね独立した内容がつぶやかれていると言える。

取得できた文字のパタンから文字の分布が表す平均情報量を調べると、理想的にランダムな分布に対する実際の分布の平均情報量は、カナに開く前で 64.8%、後で 87.7%であった。ランダムに近い分布であるほど、検索時に特徴を元にした検索範囲の枝切りを回避できる。よって、音声に合わせて問題文をカナで記述することで、検索による文の探索が困難になることが確認できた。

3.2 問題文長と問題に適した素材文章

問題文として利用するには、人間が自然な文とワードサラダ文を識別可能な情報が含まれる十分な文字数が必要となる。これは実際に実験を通した調査が必要であるため、本稿では 20 文字以上を暫定値とする。この基準において問題文として利用できるツイートは、本データでは約 54%であり、作問に必要な十分な文章量を期待できる。

問題文長と検索を用いた攻撃に対する安全性の関係について検討する。提案方式では元文章を改変するため、検索語を形態素に分割し、N の小さい形態素 N-gram で構築されたデータベースで検索し、その AND した出力を用いるのでは正しい結果は得られない。解決方法としては、問題文長に合わせた N-gram データベースの利用が考えられる。本データで 20 文字に相当する形態素数は約 13 なので、13-gram データベースが必要になるが、このサイズは非常に大きい [3] 。つまり、問題回答集や検索用のデータベースの作成、それを利用したプログラムをメモリ内で納めて動作させるのはいずれも困難であり攻撃を妨げる。

4. 回答形式とユーザビリティ

1 問の総当り攻撃成功確率 P の攻撃者が、n 問試行し k 問正解する場合を考える。このときの総当り攻撃成功率は、式(2)で表される。式(2)の値 1%未満を、要求仕様とする。

$$\Sigma_{i=k}^{n}\ {}_{n}C_{i}P^{i}(1-P)^{n-i} \qquad (2)$$

択一回答の場合、4 択であれば P=1/4 となり、(n,k)=(4,4) で仕様を満たす最小の n を取る。一方、複数文を表示し、それぞれについて文章の特徴 あり／なし を回答させる場合、4 文であれば P=1/16 となり (n,k)=(2,2) でよい。これは、総当り攻撃に対する安全性を保ちつつ、利用者が読む必要のある文を4 × 4 から4 × 2 に削減することで、ユーザビリティを向上させている。文章量の削減は、音声利用者に対して、特に効果的である。

参考文献

[1] Jeffrey P. Bigham and Anna C. Cavender, "Evaluating Existing Audio CAPTCHAs and an Interface Optimized for Non-Visual Use", SIGCHI 2009, pages 1829-1838, ACM.

[2] 山口 通智, 中田 亨, "人間ロボット判別テストのバリアフリー化のための言語的作問技法", 情報処理学会研究報告, CSEC, 2013(30):1-8.

[3] 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, データベース・システム研究会報告 2009, 148:1-8.

[4] 奥野陽, 颯々野学, "大規模日本語ブログコーパスにおける言語モデルの構築と評価", NLP2011, C4-5, pages 955-958.