様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　５ 年　 ６ 月 １５ 日現在

機関番号：３２６８９
研究種目：挑戦的研究（萌芽）
研究期間：2020～2022
課題番号：２０Ｋ２１７９７

研究課題名（和文）文字セキュリティの開拓

研究課題名（英文）Research on Security of Characters

研究代表者
森　達哉（Tatsuya, Mori）

早稲田大学・理工学術院・教授

研究者番号：６０７０８５５１

交付決定額（研究期間全体）：（直接経費）　　5,000,000円

研究成果の概要（和文）：本研究課題は、ラテン文字のaとキリル文字のａのように、形状が似ていて異なる符号が割り当てられている文字のペア「ホモグリフ」に着目した研究である。多くの人間はホモグリフに気づかない一方、自然言語処理ソフトウェアはその違いを反映するため、固有のセキュリティリスクがある。研究の結果、機械翻訳システムのホモグリフ処理に課題があること、およびニューラルネットワークだけでなく、テキストの前処理も結果に大きく影響を及ぼすことを明らかにした。また、本研究課題の応用として、人間には可読なテキストを表示するが、実際には異なる文字符号のデータをブラウザが処理することで、テキストの著作権保護を実現する方式を開発した。

研究成果の学術的意義や社会的意義
本研究はホモグリフに関連するセキュリティ課題を探求した。その応用範囲は広範であるため、波及的効果が見込める。また、文字はブラウザやアプリケーションなど様々な場面で扱われ、最近注目を集める大規模言語モデルでも重要な役割を果たす。本研究の成果は、文字を扱うアプリケーションのセキュリティリスクを低減し、より安全なデジタル環境を提供するために必要な新たな手段を示している。以上のことから、本研究はその学術的価値に加え、社会的意義も大いに有する。

研究成果の概要（英文）：This research project focuses on "homoglyphs," pairs of characters such as the Latin "a" and the Cyrillic "а" that look similar but are assigned different code points. While many people overlook these homoglyphs, natural language processing software reflects their differences, creating unique security risks. The results of this study highlight the challenges of dealing with homoglyphs in machine translation systems and show that not only the neural network, but also the preprocessing of the text significantly affects the results. Furthermore, as an application of this research, we developed a method for copyright protection of text that displays human-readable text but processes data with different character codes in the browser, effectively masking the original content.

研究分野：情報セキュリティ

キーワード：ホモグリフ　セキュリティ　機械学習　認知

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

自然言語処理は、文書分類、情報検索、感情分析、テキストマイニング、スペル訂正、クエリ提示、機械翻訳、対話システム等、多岐にわたる技術を開発し、私たちの日常生活に深く関わっている。これらの技術を適切に動かすためには、文字を符号化した符号化文字集合の導入が欠かせない。今日、広く使われている符号化文字集合の一つである Unicode は、Web サイトの 94%以上で利用されており、その最新版は合計 150 種類以上の言語に対する約 14 万の文字符号を収録している。
符号化文字集合には、ホモグリフ（外形的に類似した字のペア）に起因する問題が存在する。例えば、ラテン文字の 'a' とキリル文字の 'а' のように、形状が似ていて異なる符号が割り当てられている文字、すなわちホモグリフが存在している。漢字の「卜」とカタカナの「ト」も同じような例である。
一般に、ホモグリフは、人間の視覚には区別が難しく、多くの人間が誤認する可能性が高い。しかし、自然言語処理ソフトウェアはこれらの差異を正確に認識し、処理を行う。その結果、人間の認識と機械処理の結果との間にギャップが生じ、新たなセキュリティリスクとなる可能性がある。例えば、攻撃者がホモグリフを悪用し、テキストデータに加えることで、意図的に人間の認知と自然言語処理の結果との間にギャップを生じさせるという脅威が存在する。自然言語処理技術がますます日常生活に浸透する現在、このような脅威を理解し、低減する方法を開発することは重要な課題である。

２．研究の目的

本研究の目的は、自然言語処理におけるホモグリフ攻撃の脅威分析と対策手段の開発である。まず、ホモグリフ攻撃が及ぼす具体的な影響とその脅威度を詳細に理解することで、実際のアプリケーションへの影響を把握する。続いて、この脅威に対抗するための具体的な防御策を開発する。これらを達成することで、安全かつ信頼性の高い自然言語処理の実現を目指す。

３．研究の方法

(1) 攻撃研究による脅威分析

自然言語処理の応用として機械翻訳、剽窃チェック、感情分析、規約違反文書フィルタ等を対象とする。これらの応用例は、悪意を持つ攻撃者が剽窃検出の回避、評判情報の改ざん、フィルタの回避といった目的でホモグリフ攻撃を行う可能性がある。この研究では、英語と日本語を対象とする。

攻撃の脅威評価では、ホモグリフをテキストデータに追加し、それが各アプリケーションの出力にどのように影響するかを実験的に明らかにする。その際、使用するホモグリフの種類、数、追加する位置を調整パラメータとする。そして、現時点で最も優れた性能を示す実装（ホワイトボックス）と商用製品やサービス（ブラックボックス）を対象に評価を行う。

(2) 人間の認知評価

ホモグリフが追加されたテキストを人間がどの程度認識できるかを評価する。具体的には、ホモグリフが追加されたテキストを見たときに、何か異常を感じるかどうかを調査する。この調査では、ホモグリフの種類や量、テキストの表示サイズ、フォント種類、利用端末など、ユーザーインターフェースに起因する要素と、言語に対する経験や知識など、人間に起因する要素が認知にどのように影響するかを明らかにする。

(3) 対策技術の確立

ホモグリフ攻撃への対策技術を確立する。これは、自然言語処理がテキストデータに適用される前に行うサニタイジング処理(データを無害化するための前処理)として実現する。具体的には、各自然言語処理応用の処理に依存しない、汎用的な対策技術の確立を目指す。Unicode 12.1 に含まれる約 13.7 万の符号化文字集合を対象とし、ホモグリフとして使われる可能性のある類似文字を網羅的に収集し、データベースとしてまとめる。このデータベースにより、入力となるテキストデータにホモグリフが含まれているかを容易に判定できる。

さらに、ホモグリフを含む単語に対するコンテキスト分析を行い、本来の意図と一致すると予想

される元の単語を推定する技術を開発する。その主要なアイディアは、対象となるホモグリフを含む単語に類似した通常の単語の候補を列挙し、前後の単語から計算される遷移確率を基に、本来の意図と一致する最も可能性の高い単語を抽出することである。また、統計的アプローチとともに、最も攻撃が成功しやすいパターンを明らかにし、そのようなパターンを効果的に検出するルールベースのアプローチも試みる。

４．研究成果

(1) ニューラル翻訳に対するホモグリフ攻撃の脅威評価

本テーマは、ニューラル機械翻訳（NMT）システムへのホモグリフを用いた敵対的攻撃の影響を評価したものである。敵対的攻撃は、元の入力に微細な変化（ホモグリフを利用）を加えて生成され、翻訳結果を意図的に操作する手法である。本研究の目的は、この攻撃が NMT システムの信頼性と精度に与える潜在的な影響を明らかにすることである。図 1、図 2 に具体例を示す。

図 1　翻訳機に対する敵対的な攻撃の例，「ソフトウェアは第三者の権利を侵害した」という文の誤翻訳を誘発．Google 翻訳で確認．

a	ɑ	n	ո
b	Ь	r	ɾ
c	с	w	ԝ

図 2　ホモグリフの例，灰色が正しいアルファベット

代表的な NMT システムとして Google 翻訳を採用し、一般的な文に対して敵対的入力攻撃を実施した。具体的な攻撃手法としては、元の文に微細な摂動を加えて翻訳結果の意味やニュアンスを変化させる方法を採用した。元の英語文のうち，摂動が加えられる単語を「摂動単語」、摂動を加えた単語を「敵対的単語」と呼称する。敵対的単語生成アルゴリズムを右に示す。

Algorithm 1 敵対的単語生成アルゴリズム

Input: 摂動単語 w_p，閾値 $border$，元の文から摂動単語を抜いた $sentence = [w_1, ..., w_{p-1}, w_{p+1}, ..., w_n]$
Output: 敵対的単語 w_{adv}

function Word_perturbation(w_p, $border$, $sentence$)
  for w_{adv} in Brute force w_p **do**
    Search match_word_list W_l that matches a regular expression
    for word w_i in W_l **do**
      Compute cosine-similarity C_{w_i} between w_i and $sentence$
    end for
    Sort(C_w) according to C_{w_i}
    for C_{w_i} in C_w **do**
      Compute and add weight cosine-distance C_d between w_{adv} and w_p
    end for
    if $border < C_d$ **then**
      return w_{adv}
    end if
  end for
  return None
end function

また、攻撃の結果として翻訳に生じた変化を下記の表のように分類する。

表 1　翻訳結果のラベルの意味 (下線がホモグリフに変更した文字)

ラベル	意味	入力例	出力例
normal	原文と意味がほとんど変わらない	We had dinner together at a curry restaurant.	カレー屋で一緒に夕食を食べました。
odd	摂動を加えた単語の文字が出力に現れる，又は文法が壊れている	He will bake a cake from scratch.	彼はscratch からケーキを焼きます。
semi-odd	摂動を加えた単語がカタカナとなって出力に現れる	He heard that for the first time.	彼は最初のテムのためにそれを初めて聞いた。
remove	摂動を加えた単語が出力から消え，意味が少し変化した	Their eye bf is slightly different.	彼らの色はわずかに異なります。
replace	摂動を加えた単語が別の意味に置き換わった，または文の意味が大きく変化した	Two of them have a conversation.	それらのうちの 2 つは会議を持っています。

敵対的入力により、翻訳文の意味やニュアンスは、微細な摂動によって変化し、意味が変更された率は 55%にも及ぶことを明らかにした。さらに、本手法の成功率は、従来のテキスト分類アプリケーションを対象とした攻撃手法に比べて高いことを明らかにした。

手法	normal	odd	semi-odd	remove	replace	意味変更率	編集距離
ランダム insert	73.0%	3.0%	1.0%	6.0%	13.0%	20.0%	1.00
ランダム delete	60.0%	8.0%	9.0%	4.0%	19.0%	32.0%	1.00
ランダム swap	58.0%	18.0%	10.0%	4.0%	10.0%	24.0%	2.00
アルゴリズム insert	67.0%	7.0%	3.0%	10.0%	13.0%	26.0%	1.00
アルゴリズム delete	62.6%	10.1%	6.1%	6.1%	15.2%	27.3%	1.00
アルゴリズム swap	48.3%	11.5%	8.0%	5.7%	26.4%	40.2%	2.00
提案手法	36.0%	9.0%	4.0%	11.0%	**40.0%**	**55.0%**	2.17

(2) 実世界における機械翻訳サービスに対するホモグリフ攻撃の評価

オンラインで使用される機械翻訳システムが、ホモグリフ攻撃に対してどのような脆弱性を持つのかを調査した。対象としたシステムは、Google 翻訳、Bing 翻訳ツール、Systran Translate、DeepL、Excite 翻訳、みらい翻訳、Weblio 翻訳、CROSS-transer の 8 つの機械翻訳システムである。以下は評価結果の概要である。

翻訳機	normal	copy	remove	fix	odd	confuse	mark	change	ゼロ幅文字攻撃
Google [1]	1	8	1	6	0	0	0	2	不可
Bing [3]	1	6	3	2	1	0	1	4	可
Systran [4]	1	17	0	8	0	0	0	2	可
DeepL [5]	4	8	1	2	0	8	0	1	不可
Excite [2]	0	17	0	0	0	0	0	0	可
みらい翻訳 [6]	5	0	1	4	2	0	0	7	不可
Weblio [7]	0	17	0	4	0	0	0	0	可
CROSS-transer [8]	3	0	3	5	0	0	0	6	可

前処理の影響を把握するために、新たに下記の分類を追加している。

copy	一部をホモグリフに置き換えた単語が出力にそのまま表れた	The sound is unpleasant.	音がunpleasant。(Bing 翻訳で確認)
confuse	出力側に，文章とは関係ない文字列が多く見られた	The answer is incorrect.	答えは、acederealedLu$_1$3A5*reasonablencorrect* です。(*DeepL* で確認)
mark	ホモグリフに置き換えた単語が，出力で別の記号となって表示された	This is not a pen.	これはペンの「」です。(Bing 翻訳で確認)
change	ホモグリフに置き換えた単語が，出力で別の意味となって表れた	This is a pen.	これはラップです。(みらい翻訳で確認)

結論として、多くのケースは copy に相当する処理をしている一方で、copy が観測されないシステムが存在、それらのシステムは、NMT への入力前に文字列に前処理を適用していることが推察される。特にみらい翻訳オンラインでは、キリル文字を採用した際に，出力が odd ラベルとなり、入力言語と異なる言語の文字が入力された際、異なる言語の文字を何らかの入力言語の文字に変換する前処理を行っていること、さらに該当する異なる言語の文字を英語で読んだ際の頭文字に影響されている可能性が高いことを明らかにした。このように、実際の機械翻訳システムにおいてもホモグリフ攻撃に対する脆弱性が存在すること、ならびにその脆弱性は NMT のみならず、前処理の方式にも起因することを明らかにした。

(3) 機械処理と人間認知のギャップを利用した著作権保護方式

本研究課題の技術的本質は、人間が目にするテキストの外形的特徴と、そのテキストを機械が処理する際の符号化文字の符号が必ずしも一致していないことにある。上述した(1)、(2)の研究テーマを進める中で、このようなギャップを攻撃として捉えるだけではなく、有用な目的に活用することの重要性を認識するに至り、当初の計画にはなかったテーマに取り組んだ。
基本的なアイディアは、ブラウザで用いられる Web フォントを用い、テキストコンテンツを独自のルールでエンコードすることにより、人間が見る文字と、機械が処理する文字符号との対応関係をランダム化することである。このアイディアを用いた、ユーザによるコンテンツの取得を防止するコピープロテクション方式を提案し、その有効性を評価した。ユニークな評価として、コンピュータサイエンスやセキュリティの知識がある実験参加者 11 名を対象としたユーザスタディで、コピーガードの有効性を評価した。参加者は与えられた時間の中で、提案手法を用いたウェブサイトから、元のテキストデータを取得できるかを評価した。この結果、提案手法は高い攻撃耐性があることを確認した。

(4) DNS ルートゾーンにおける日本語ルール生成の設定

ホモグリフを用いた攻撃に対する対策の一貫として、日本語を用いたトップレベルドメインにおける利用可能文字を制定するワーキンググループに協力した。漢数字の「二」とカタカナの「ニ」などのホモグラフに対するユーザの認知を評価し、日本語が母国語であるか否かによらず、誤認が生じやすいことを明らかにした。結果は ICANN の Root Zone LGR Project に提案・承認され、全世界のトップレベルドメイン空間における普遍的なルールとして採用された。
J-LGR Proposal v0.15, Appendix B: RESEARCH PAPER: SURVEY ON THE USER PERCEPTION OF THE HOMOGRAPHIC CHARACTER SET SPECIFIED BY JGP

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Kawaoka Ryo、Chiba Daiki、Watanabe Takuya、Akiyama Mitsuaki、Mori Tatsuya	４．巻 12671
２．論文標題 A First Look at COVID-19 Domain Names: Origin and Implications	５．発行年 2021年
３．雑誌名 Lecture Notes in Computer Science book series	６．最初と最後の頁 39〜53
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/978-3-030-72582-2_3	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 河岡諒，千葉大紀，渡邉卓弥，秋山満昭，鈴木宏彰，森達哉
２．発表標題 東京2020オリンピック公式ドメイン名に対する類似ドメイン名の実態調査
３．学会等名 コンピュータセキュリティシンポジウム2021
４．発表年 2021年

１．発表者名 坂本岳史、森達哉
２．発表標題 オンライン機械翻訳システムに対するホモグリフ攻撃の脆弱性調査
３．学会等名 情報通信システムセキュリティ研究会（ICSS）
４．発表年 2021年

１．発表者名 坂本岳史，森達哉
２．発表標題 ニューラル機械翻訳システムに対する敵対的攻撃
３．学会等名 情報通信システムセキュリティ研究会（ICSS）
４．発表年 2020年

１．発表者名
野本一輝，森達哉

２．発表標題
Fonty: テキストのコピーを防止する文字交換方式の提案と評価

３．学会等名
情報通信システムセキュリティ研究会（ICSS）

４．発表年
2023年

〔図書〕　計0件

〔産業財産権〕

〔その他〕
Proposal for a Japanese Script Root Zone LGR
Appendix B: RESEARCH PAPER: SURVEY ON THE USER PERCEPTION OF THE HOMOGRAPHIC CHARACTER SET SPECIFIED BY JGP（ICANN 共同研究）

https://www.icann.org/en/system/files/files/proposal-japanese-lgr-20dec21-en.pdf
https://j-gp.jp/%E6%97%A5%E6%9C%AC%E8%AA%9ELGR%E6%8F%90%E6%A1%88%E6%9B%B8

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
米国	ICANN			