改行位置の制御によるテキストステガノグラフィの提案
— プレーンテキスト及びXMLへの情報埋込 —

滝澤修[*1]　山村明弘[*1]　中川裕志[*2]　松本勉[*3]　村瀬一郎[*4]　牧野京子[*4]　井上信吾[*4]　大野浩之[*1]

*1 総務省通信総合研究所　*2 東京大学　*3 横浜国立大学　*4 株式会社三菱総合研究所

1．はじめに

計算機ネットワークの利用拡大に伴い、ネットワーク上で情報を安全に伝送する情報セキュリティ技術が重要になってきている[2]。情報セキュリティ技術の一つである暗号は、攻撃者に解読できないように、あるいは改ざんされたらその検出ができるように、情報を加工・復元する技術であり、情報が埋め込まれていることを隠すことは必ずしも目的としていない。それに対して、情報が埋め込まれていること自体を隠す技術であるステガノグラフィ（深層暗号）は、情報伝送に際してのカムフラージュ手段、あるいは画像や音楽などの著作物に著作権情報や配布先情報を埋め込む手段として注目を集めている(注1)。ステガノグラフィは、埋め込み媒体が持つ情報の冗長性を利用して別の情報を埋め込むものであり、画像や音響信号などに対しては、人間に識別できるような劣化をきたすことなく比較的実現しやすい。それに対して自然言語テキスト（キャラクタコード列）を埋め込み媒体とした場合、文字コードには1ビットの冗長性も無い[3]ため、情報を埋め込むことで1ビットでも改変されると正しい文字として再生されず、見た目の変質をきたすのみならず、秘匿情報の存在が容易に露見してしまう。そのためこれまでテキストステガノグラフィとしては、著作物の権利主張を目的とするよりは、情報伝送に際してのカムフラージュ手段とするものが多く、テキストの内容保持には重きを置かず、自然言語としての自然性を保つ方法にもっぱら重点を置いた研究[4]が多かった(注2)。

本稿では、著作物の権利主張手段とすることを目指し、日本語文を対象として、改行位置を制御することにより情報を埋め込むテキストステガノグラフィとして、2つの手法を提案する。またXMLを対象としたステガノグラフィについて考察する。

(注1) 情報を秘匿することを目的とする技術全般を指して「情報ハイディング」(Information Hiding)という用語が用いられることが多い。情報ハイディングは、情報の存在を秘匿する技術であるステガノグラフィと、通信の秘匿化（経路の秘匿化、秘密通信路など）とを包括した広い概念である[1]。本稿では、通信の秘匿化は対象とせず、情報の埋め込み方に限定した議論をしているので、ステガノグラフィという用語を用いることにする。

(注2) 従来、テキストを対象とするステガノグラフィのほとんどは、テキストを印字画像として扱い、字間、行間、文字の大小・回転などの制御によって情報を埋め込む手法であった。これらはテキストをキャラクタコード列として扱うものではないため、本稿でいうテキストステガノグラフィとは異なる。

2．改行位置の制御によるテキストステガノグラフィの特長

本節では、先行研究において提案されている主な手法を概観し、提案する手法と比較検討する。

SNOW[5]は、英文を埋め込み媒体とし、視認しにくい複数の空白文字を行末に挿入することにより情報を埋め込む手法である。埋め込み情報はハフマン符号化により圧縮し、独自の手法[6]により暗号化した後、行末に0～7個の空白を挿入することによって1行当たり3ビットの情報を埋め込む。本手法では、空白文字を無視あるいは空白として表示する出力系においては、見た目の文書の変質は起きない。しかし行末に不審な文字コード（空白文字）が多数存在していることが機械的にすぐ見つけられる可能性があるため、作為を発見されやすい懸念がある。また上記の出力系による印字出力では埋め込み情報が消えてしまい、また無効化攻撃(注3)にも弱い。

松本ら[7]は、法執行機関による合法的な暗号傍受を逃れる手段として、より深層の暗号通信を使う可能性を検討した中で、多重暗号化法の具体例として、英文のLaTeXソースを埋め込み媒体とし、本文の各行の単語の個数を加減することにより、情報を埋め込む方式を検討している。この方式は、一般的なLaTeXの出力系において、コンパイル後の表示文書が埋め込む前後で全く変わらない性質を利用しており、計算機プログラムへの情報埋め込み手法の一種と捉えることができる。但しLaTeXソースを自然言語テキストとみなした場合には、本稿で提案する方法2を、膠着言語ではない英語に適用した場合に相当する方式であるといえる。

FinPri.txt[8]は、日本語文を埋め込み媒体とし、意味的に近い単語で置き換える、いわば語彙の冗長性を利用することによって、文章の意図を大きく変えないで情報を埋め込む手法である。文書の意図を保存することに主眼を置いた本手法は、マニュアル文などの技術文書には有効である。またどの単語をどのように置き換えることで情報が埋め込まれているかを攻撃者が発見しにくいため、無効化攻撃に比較的強い特長があ

(注3) ステガノグラフィを脅かす攻撃として、埋め込まれた情報を破壊・消去してしまう「無効化攻撃」、情報が埋め込まれている事実を見つけ出す「検出攻撃」、埋め込まれた情報の内容を解読する「抽出攻撃」の主に3つがある。画像ステガノグラフィの多くは、画像サイズを変えたりファイル形式を変換するなどの一般的処理により、見た目の品質を損なうことなく、埋め込まれた情報が破壊・消去されてしまうことが多い[3]。これは結果的に無効化攻撃となる。無効化攻撃に強いステガノグラフィを実現するためには、無効化をすれば埋め込み媒体自体が損なわれるような埋め込み方法を講じる必要があるとされる。

る。しかし、文芸作品や契約文など微妙なニュアンスが重要視される文書では、意味的に近い単語であっても置き換えることによって文書に変質劣化をきたす可能性がある。また語彙の改変は著作権にかかわることであるため、著作権者自らによる処理以外は法律上の制約を受ける恐れがある。

これらの手法に対し、本稿で提案する手法は、英語のようなハイフネーションが不要な膠着言語としての日本語の性質を生かし、単語を切断するような改行を積極的に行うことで、プレーンテキストの行末の僅かな凸凹に情報を埋め込むものであり、以下の特徴をもつ。

(1) 空白文字のような無意味なコードを挿入せず、テキストにとって必須な改行コードのみを利用して埋め込む手法のため、不自然さや作為の発見されやすさが少ない。
(2) 改行コードを無視する出力系ではない場合、印字出力にも埋め込み情報が残るため、電子的な流通だけでなく印字出力としての流通にも有効である。
(3) 単語を切断する改行を許容する方式のため、一行文字数のぶれを小さく抑えることができ、改行位置に情報が隠されていることは見抜かれにくい。均等割付された印字出力としての使用を想定すれば一層見抜かれにくいが、均等割付しない場合でも、一行の長さが不揃いな文書は珍しくないため、機械的な手段による検出攻撃は困難と思われる。
(4) 単語は全く置き換えないので、文書の変質劣化はきたさない(注4)。そのため文芸作品や契約文などにも適用でき、著作物の権利主張手段として適している。
(5) 意図的・非意図的な文書の整形による無効化攻撃には弱い。また結託攻撃(注5)にも弱い。但し一行文字数を揃える機械的な整形はメーラなどで行われるものの、多くの場合、行末を折り返すだけであり、当初の改行コードは保存される。本手法において脅威となるのは、改行コードを削除したり改行位置をつけかえたりするような整形であるが、これは、本文と章題や箇条書きとの区別の困難さや、段落途中と段落末との区別の困難さなどから、プレーンテキストを対象としたアプリケーションにおいてはあまり行われていない。そのため、非意図的な整形による無効化については、懸念する必要は少ないものと思われる。

3. 提案する手法

3.1 概要

提案する手法は、以下の考え方に基づきアルゴリズムを考えることになる。

(1) 埋め込み媒体は、段落毎に改行コードが入ったテキストとする。これは一般のワープロ文書の形式と考えられる。
(2) 無効化・検出・抽出の3攻撃(注3)のうち、本手法は無効化攻撃には弱い。しかし少なくとも抽出攻撃(解読や、なりすまし)は回避する手立てを講じる必要がある。従って埋め込み情報は暗号化しておく必要がある。
(3) 埋め込まれている場所を攻撃者に特定されにくくするため、埋め込み媒体となる文章の中における埋め込む位置は不定とする。そのため、データの始まり/終わりを表すフラグシーケンスをデータの前後に付加する。

改行位置と埋め込み情報との対応づけについては、単語中の改行位置による方法(方法1)と、一行文字数による方法(方法2)の2つの方法を検討した。以下ではそれぞれについて述べる。

3.2 単語中の改行位置による情報埋め込み (方法1)

方法1では、形態素解析辞書の見出し単語を対象に、各単語(形態素)中の改行位置と、埋め込み情報のビットとの対応関係に基づき情報を埋め込む。例えば図1に例示するように、「する」を「す|る」と改行したら"1"などと予め定めておく("|"は改行位置)。埋め込み処理時に指定する一行基準文字数に従い、行末の近傍に来た単語を埋め込み対象とする。図1に示すように、「プログラミング」や「コミュニケーション」などの長い単語は、複数の改行位置を0,1に対応させておき、どれを選んでもいいようにしておく。こうすることで一行基準文字数から大きくかけ離れない文字数で改行できる。

0	1	
する|	す|る	(動詞-自立 サ変・スル)
プログラミング|	プログラミン|グ	(名詞-サ変接続)
プロ|グラミング	プログ|ラミング	(名詞-サ変接続)
言|語	言語|	(名詞-一般)
獲得|	獲|得	(名詞-サ変接続)
コミュニケーション|	コミュニケーショ|ン	(名詞-一般)
コミュニケ|ーション	コミュニケー|ション	(名詞-一般)
コミュ|ニケーション	コ|ミュニケーション	(名詞-一般)
役立つ|	役|立つ	(動詞-自立 五段・タ行)
と|して	として|	(助詞-格助詞-連語)
同時に|	同時|に	(副詞-一般)
こと|	こ|と	(名詞-非自立-一般)
考|え	考え|	(動詞-自立 一段)
言語|	言|語	(名詞-一般)
そこで|	そこ|で	(接続詞)
研|究	研究|	(名詞-サ変接続)

図1 形態素毎のビット対応表の例
(形態素は参考文献[9]の付属辞書に基づく)

図1の対応表を用いて情報を埋め込んだ例を図2に示す。情報を埋め込んだ単語(形態素)を下線で示している(下線は実際には非表示)。図2は均等割付をしたものであるが、一行文字数のバラツキはほとんど視認できないことがわかる。図2の例では、最初の3行は情報を埋め込んでいないダミー改行で、

(注4) 詩のように、改行位置が重要な役割を果たす文章もあるが、本研究では対象外とする。

(注5) 同一のコンテンツに異なった情報を埋め込んで作られた複数のコンテンツを照らし合わせることによって、情報の埋め込み方を探る攻撃を、結託攻撃という。

4行目から10行目までが始まりのフラグシーケンス“0111110”、11行目以降(“1011…”)が埋め込まれた情報の本体となる。

本手法は、以下の特長を持っている。

(1) 字種(ひらがな/カタカナ/漢字)による切り分けを行えば、形態素解析を使わず軽い処理が可能。

(2) 単語単位で埋め込み方を定義できるため、後述する方法2と比較して、埋め込み情報のビットと改行との対応関係の法則性を見破ることが困難であり、従って抽出攻撃(注3)に強い。

(3) 単語毎に改行位置を定義できるため、不自然な位置での改行を回避することが可能。

一方、課題としては、形態素解析処理の誤りへの対処、一文字形態素への対処などがある。

自然言語は、冗長性、文脈依存性、解釈多様性などの曖昧性を本質的に持っています。自然言語における曖昧性の存在は、言語哲学あるいは認知科学上の考察の対象としては面白いのですが、機械翻訳などの実用的な自然言語処理にとっては、性能向上を阻害する 0
困った性質といえます。なぜ人類はこれまでの進化において、プログ 1
ラミング言語のような、もっと曖昧性の少ない効率的な自然言語を獲 1
得してこなかったのでしょうか。それは、曖昧性がコミュニケー 1
ションにとって必要だからではないかと思われます。曖昧性が役 1
立つ例として、大量の意味を少ない言葉に含めたり、複数の意味を同時 1
に伝えたりできることや、特定の相手にだけ真意を伝えられること 0
、状況の変化に応じて新たな意味を容易に定義できること、などが考え 1
られます。無限の状況を有限の言葉によって表現できるのも、自然言語 0
が曖昧性を持っているがゆえに可能なのではないでしょうか。そこ 1
で、自然言語が持つ曖昧性に積極的に着目し、工学的に扱うための研究 1
は、大変重要なものです。
....

図２　方法１により情報を埋め込んだ例
(右端の数字は埋め込まれた情報(実際は非表示))

3.3　一行文字数による情報埋め込み（方法２）

方法2では、埋め込み処理時に一行基準文字数を指定し、各行の文字数と埋め込み情報のビットとを対応させる。例えば文字数が偶数なら 0、奇数なら 1 などとする。本方法を用いて情報を埋め込んだ例を図3に示す。

自然言語は、冗長性、文脈依存性、解釈多様性などの曖昧性を本質的に持っています。自然言語における曖昧性の存在は、言語哲学あるいは認知科学上の考察の対象としては面白いのですが、機械翻訳などの実用的な自然言語処理にとっては、性能向上を阻害する 0
困った性質といえます。なぜ人類はこれまでの進化において、プ 1
ログラミング言語のような、もっと曖昧性の少ない効率的な自然 1
言語を獲得してこなかったのでしょうか。それは、曖昧性がコミ 1
ュニケーションにとって必要だからではないかと思われます。曖 1
昧性が役立つ例として、大量の意味を少ない言葉に含めたり、複 1
数の意味を同時に伝えたりできることや、特定の相手にだけ真意を 0
伝えられること、状況の変化に応じて新たな意味を容易に定義で 1
きること、などが考えられます。無限の状況を有限の言葉によって 0
表現できるのも、自然言語が曖昧性を持っているがゆえに可能な 1
のではないでしょうか。そこで、自然言語が持つ曖昧性に積極的 1
に着目し、工学的に扱うための研究は大変重要なものです。
....

図３　方法２により情報を埋め込んだ例
(右端の数字は埋め込まれた情報(実際は非表示))

図3の例は、一行基準文字数を全角30文字とし、図2と同じく、最初の3行は情報を埋め込んでいないダミー改行で、4行目から10行目までが始まりのフラグシーケンス“0111110”、11行目以降(“1011…”)が埋め込まれた情報の本体となる。

本手法は方法1のような辞書照合を必要としないため、処理が速く誤処理が少ないと考えられる。反面、埋め込み方の法則性が平易なので、抽出攻撃の危険性が高い問題がある。課題としては、不自然な位置での改行の回避方法や、半角文字と全角文字とが混在した場合の文字数の計数方法などがある。

4．議論

本稿で提案した２つの手法は共に、情報を埋め込んでも本文の意味内容に全く影響しない特長がある。また、1 行につき必ず 1 ビットは情報を埋め込む方式なので、埋め込める情報量を保証できる特長がある。

両手法に共通する検討課題として、より不自然でない改行位置の制御がある。即ち、禁則処理や章題、箇条書き、固有名詞等の扱いを定義する必要がある。

Maxemchuk は、電子すかし技術が満たすべき条件として、以下の３つを提示している[10]。

(1) 埋め込まれた情報の除去が困難であること。

(2) 埋め込まれた情報を除去した場合、除去された事実がわかるようになっていること。

(3) 埋め込まれた情報を除去した場合、埋め込み媒体の質の低下をきたすようになっていること。（除去したら使いものにならなくする）

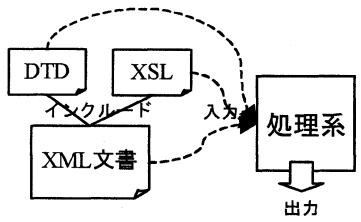
テキストステガノグラフィの場合、埋め込み媒体であるキャラクタコード列自体には1ビットの冗長性も無いため、テキストの見た目の変質をきたすことなくこれらの条件を満たすことは原理的に困難である。そのため、情報が埋め込まれていること自体を隠すことに第一義を置くものとし、いかに自然性を損なうことなく埋め込まれているかを評価の尺度とすべきである。その点で本手法は自然なものであり、上記3条件を満たさなくても支障の少ないアプリケーションに限定すれば、利用価値があるものと考えられる。

提案した手法は、単語の置き換えや、後述するマークアップ言語のタグへの埋め込みのような、テキストのキャラクタコードを細工する手法と、行間隔や文字の傾きの細工など、テキストを画像として扱う手法との中間に位置づけられ、両者の特長を併せ持っていると言える。前者は、文書をコードとして処理するデジタル的な流通への適用には不可欠な手法であるが、情報を埋め込むことが難しい。後者は印字出力としての使用が前提なので、画像ステガノグラフィの手法を適用でき、多くの情報を違和感なく埋め込める可能性があるが、デジタル的な流通には適用できず、光学的な複写の防止程度の用途しかない。それぞれの特長を生かした利用法を考える必要がある。

5．XMLに対するテキストステガノグラフィ

前節までに提案した手法は、プレーンテキストを主に対象としている。今後、電子コンテンツが普及していくと、プレーンテキストに限らず、SGMLやHTMLなどのマークアップ言語の形式で流通する場面が増えることが考えられる。その中で近年、構造化言語として、XML(eXtensible Markup Language)が注目を集めている。そこで本節では、テキストステガノグラフィの可能性の一つとして、XMLへの情報埋め込み方法について検討する。

XMLにおいては、「内容」「構造」「体裁」を別々に扱う[11]。「内容」をタグによってマークアップしたものを「XML文書」と呼び、その「構造」はDTD(Document Type Definition)と呼ばれる構造定義体によって定義される。「体裁」はスタイルシートにより記述され、CSS(Cascading Style Sheet)とXSL(eXtensible Stylesheet Language)の2種類がある。なお、XSL(またはCSS)は必須ではない。図4に、XML文書、DTD、XSL(またはCSS)の関係と、処理の概要を示す。



XML文書、DTD、XSLはそれぞれ別のファイルであり、それぞれが処理系の入力となり、処理系は3つを統合して処理し、出力を行う。

図4 XMLにおける処理の概要

XMLにおけるステガノグラフィを検討する際、XMLの利用法を想定した上で埋め込み媒体を考える必要がある。現状における利用法を2つに場合分けすると、それぞれの埋め込み媒体は以下の通りになる。

(a) XMLをデータ交換基盤として利用する場合
　一般に業界毎にDTDが決定され、日常的にはXML文書のみが流通する。その場合、情報の埋め込み者がDTDを変更する余地はないため、情報の埋め込み媒体としては、XML文書のみと想定することが妥当である。

(b) XMLをHTMLの代用として、主としてブラウザで閲覧するために用いる場合
　XML文書を作成する利用者は、DTDおよびXSL(またはCSS)を個々のXML文書毎に生成することが考えられる。そこで情報の埋め込み媒体としては、XML文書に加えて、DTDおよびXSL(またはCSS)も対象とすることが妥当である。

以上の考察を踏まえると、情報の埋め込み方法として、以下のものが考えられる。

(1) XML1.0[12]で定められているXMLの仕様の曖昧さを利用して情報を埋め込む方法
(2) DTDやXSL(またはCSS)におけるタグ定義に曖昧さを設けて情報を埋め込む方法

6．まとめ

本稿では、日本語文を対象として、改行位置を制御することにより、文章の意味を全く変えることなく情報を埋め込むテキストステガノグラフィとして、2つの手法を提案した。また、XMLを対象としたステガノグラフィについて基礎的検討を行った。

本稿では手法の概念提案にとどまり、インプリメントによる実証および問題点の抽出までは至らなかったが、今後、今回提案した方法の他にも各種方法を提案し、実装した上で不自然さの評価や気づかれにくさの評価を行う。また、XMLを対象としたステガノグラフィについても、引き続き検討する。

さらに、テキストステガノグラフィ全般について、その位置づけや分類、利用方法について整理する予定である。

【謝辞】

本研究の実施にあたって、定期的な意見交換により有益な助言をいただいた横浜国立大学松本研究室および(株)三菱総合研究所の各位に感謝する。

【参考文献】

[1] 情報処理振興事業協会,「インフォメーションハイディングの技術調査」報告書, http://www.ipa.go.jp/security/fy10/contents/crypto/report/Information-Hiding.htm, 平成10年2月.
[2] 辻井重男,「暗号と情報社会」,文藝春秋,1999.
[3] 松井甲子雄,「電子透かしの基礎」,森北出版,1998.
[4] 中川,木村,三瓶,松本,「辞書変換法に基づく日本語テキストへの情報ハイディング」,情処論, Vol.41, No.8, pp.2272-2279, 2000.
[5] "The SNOW Home Page", http://www.darkside.com.au/snow/, Feb. 2001.
[6] M. Kwan, "The Design of the ICE Encryption Algorithm", proceedings of Fast Software Encryption - Fourth International Workshop, Haifa, Israel, Springer-Verlag, pp. 69-82, 1997.
[7] 松本, 糸山, "Lawful Accessの無効化を狙う暗号通信の検出は容易か？", 信学技報 ISEC96-79, pp.159-164, 1997年3月.
[8] 松本, 中川, 村瀬, "ネットワーク向けインフォメーションハイディング技術開発 テキスト用フィンガープリンティング方式 FinPri.txt の開発", 情報処理振興事業協会 次世代デジタル応用基盤技術開発事業 先端的情報化推進基盤整備事業 論文集, pp.97-104, 2000年6月.
[9] "日本語形態素解析システム茶筌 version 2.0 for Windows", 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座(松本研究室), 1999.
[10] N.F.Maxemchuk, "Electronic Document Distribution", AT&T Tech.J., Vol.73, No.5, pp.73-80, 1994.
[11] 野田俊太郎, "XML入門", http://www.utj.co.jp/xml/beg/guide/xml_1.html, 2001年2月.
[12] "Extensible Markup Language (XML) 1.0 (Second Edition)", http://www.w3.org/TR/REC-xml, Feb.2001.