

4. ドキュメントへの インフォメーションハイディング

中川裕志

東京大学
nakagawa@dl.itc.u-tokyo.ac.jp

滝澤 修

通信総合研究所
taki@crl.go.jp

井上信吾

(株) 三菱総合研究所
sinoue@mri.co.jp

■ドキュメントへの情報ハイディング とは

カバーデータとしてドキュメント（文書）を用いる情報ハイディングは、第三者が気づかない作為をドキュメントに施すことによって情報を埋め込み、正当な者のみがそのドキュメントから秘匿情報を抽出できるようにすることを指す。

有史以来使われてきた古典的な情報ハイディングは、もともとドキュメントを媒体とするものが多くを占めていた。現代においては、第三者による傍受や検閲等の脅威に対抗することなどを想定したステガノグラフィ（秘密通信路、あるいは秘匿通信）としての応用がまず考えられる。ドキュメントにおけるステガノグラフィは、第三者が通常の通信とみなすデータに実は第三者の目を逃れる秘密の情報が埋め込まれているモデルである。一方、大規模災害などの非常時において、誰もが目にできる文書、たとえば災害情報掲示板の中に、パニック等の回避のため一般大衆に解読されることを避けつつ、しかるべき人、たとえば消防士だけが解読できるクリティカルな情報（混雑していない避難路など）を埋め込むような応用も考えられる。そのほかに、ステガノグラフィの一種として、視覚復号型秘密分散法をドキュメントへ適用し、複数のドキュメントを重ね合わせた時にだけ秘匿情報が現れるテキスト秘密分散の開発も試みられている¹⁾。

ステガノグラフィとともに、ドキュメントへの情報ハイディングの有力な応用分野と考えられているのが、電子的コンテンツに対して著作権情報やフィンガープリントを埋め込む電子透かしである。これは、コンテンツを正当に入手した人や組織を特定できる情報などをコンテンツに埋め込んでおくもので、不正な2次配布をした

場合に流出元を特定できることによって、海賊版の流布に対する抑止効果が期待できる。

ドキュメントへの情報ハイディングにおいて考慮しなければならないのは、カバーテキスト（カバーデータとしてのテキスト）の改変をどれくらい許容するかである。小説などカバーテキストそのものが著作物である場合、改変はまったく許容できないと言っていい。一方、著作権を主張する主な対象がソフトウェアや画像あるいはビデオなどであり、その付属物であるドキュメントに著作権情報を埋め込む利用法の場合、たとえば、ソフトウェアならマニュアルや使用許諾書のような添付文書をカバーテキストとして情報を埋め込む場合には、ステゴテキスト（ステゴデータとしてのテキスト）はカバーテキストの意味が保存されていればよく、したがって文面に若干の変更が加えられても許容できる場合がある。さらに、秘匿されている情報に重点が置かれ、ステゴテキストはカムフラージュに過ぎない場合、機械による自動的な検閲を逃れる目的であれば、ステゴテキストは意味のある文書になっていなくても構文的に正しくさえあればいい場合もある。

情報ハイディングは、カバーデータの冗長性を利用して情報を埋め込む技術であるので、ドキュメントのどの側面の冗長性を利用するかによって、いくつかの方式に分類できる。分類の際、情報秘匿のための作為がハードコピー上（あるいはディスプレイの画面、以下同じ）に残る方式と、残らない方式の2つに大別して考えると理解しやすい。ハードコピー上に残るか残らないかの違いは出力系に依存するので、必ずしも厳密な分類とはいえないが、説明上は都合がいいので、以下では一般的な出力系を想定して両方式を概観する。

■作為がハードコピー上に残る情報ハイディング

ハードコピー上に残る方式は、作為が目視確認できるはずだが気づきにくいことを利用するもので、電子データとしてだけでなくハードコピーとしての流通にも使える特長がある一方、見破られないように作為の仕方を工夫する必要がある。この方式は、気づきにくさの原理によって、さらに以下の2つのタイプに細分類できる。

◇作為が見えにくいことを利用するタイプ

カバーテキストとステゴテキストとを並べて目視比較しても見分けられない程度の、微小な作為を施すことによって、見破れることなく情報を埋め込むことを目指す。その実現方法として、文書レイアウトへの作為が考えられる。これは、ポストスクリプト機能等を活用して、文書レイアウトに微小な作為を施し、ハードコピーとして印刷されたステゴテキストをスキャナで読み取って秘匿情報を抽出する手順が基本となる。文字情報そのものは埋め込みと抽出の両場面で重要ではなく、ドキュメントの画像としての情報の差異を利用する。そのため、これは画像への情報ハイディングの一特殊形とみなすこともできる。ハードコピーとして利用する場合、複写を繰り返して画像が劣化することにより、秘匿情報も劣化消失することが、このタイプの弱点といえる。ハードコピーを介さずデータ内に秘匿した情報を電子データのまま受け取って抽出するモデルもあり得るが、その場合はそもそもレイアウトに作為を施す必要はなく、したがって後述するXMLやLaTeX文書への情報ハイディングなどと同類とみなせる。

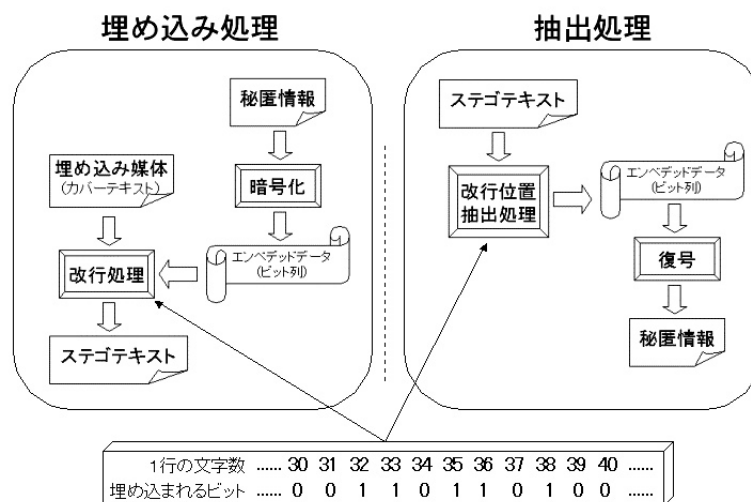
レイアウトへの作為の施し方としては、行間隔あるいは語間隔の拡大縮小や、文字幅の拡大縮小あるいは文字の回転などが提案されている。たとえば行間隔の標準画素数を定めておき、ビット“1”を埋め込むと間隔が拡大し、“0”を埋め込むと狭まるとする方法である。したがって秘匿情報の抽出性能はスキャナの読み取り解像度に依存することになるので、拡大縮小の程度を小さくすればより作為に気づかれにくいだが、半面、抽出エラーも増加することになる。どの作為が気づかれにくいかは言語に依存し、たとえば英語などの欧文では、語間隔の拡大縮小が有利で、日本語のような表意文字を多く使う膠着言語（単語間にスペースが挿入されない言語）では、フォントの拡大縮小あるいは回転を利用した作為が有利とされている²⁾。また、秘匿情報の抽出に際して、原本であるカバーテキストとの比較照合を必要とする手法としない手法とがある。レイアウトへの作為を用いる各種手法については、文献3)に多数紹介されているので、参照されたい。

レイアウトへの作為以外に、ドキュメントの周辺や罫線などにきわめて小さい文字や記号を隠しておく方式も、このタイプに属する。また筆跡の座標や筆圧に作為を施して情報を秘匿する手書きステガノグラフィ⁴⁾も、ドキュメントへの情報ハイディングとみなすならば、このタイプに属すると言えよう。

◇作為が自然なため気づかないことを利用するタイプ

デジタルドキュメントは基本的に、文字列とレイアウト情報とからなる。文字はそれ自体が意味の一部を成すため、デジタル情報としての文字へ無配慮に作為を行うと、わずかな作為であっても文字化けなどを起こし、意味にまで波及することでドキュメントの品質が大きく損なわれ、また作為が露見する恐れが高まる。そのため、ドキュメントへの情報ハイディングは従来、前述した文書レイアウトへの作為に分類される手法が多く提案されてきた。しかし電子メールのようにレイアウト情報を持たないプレーンテキストに情報を埋め込む場合は、文字への作為のみに頼る必要がある。この場合、作為の見えにくさによるカムフラージュは断念し、ステゴテキストだけを観察した場合に不自然さに気づかれにくいことでよしとする戦略をとる。この方式ではカバーテキストとステゴテキストとを並べて目視比較すると差異が歴然とするため、カバーテキストを公開しない利用モデルが想定される。本タイプは作為がかなり大きいため、ハードコピーとして利用する場合に複写を繰り返しても秘匿情報が劣化消失しにくい特長がある。

文字への作為において、ドキュメントの変質を避けるためには、単語の置き換えなど自然言語処理を応用した方法と、文面に影響しない文字あるいは文字コードを挿入する方法とが考えられる。前者の手法については、いくつかの方法論的な種類があるので、「自然言語処理を応用した情報ハイディング」の章で改めて詳述する。後者の手法としては、ドキュメントの改行する場所をコントロールすることにより情報を埋め込む方式が提案されている⁵⁾。この方式は、日本語のように、改行する位置が比較的自由的な膠着言語を対象としている。この方式による処理の流れを図-1に示す。処理においてはまず、各行の文字数と埋め込みビットとの対応表を定義しておく。そして、ドキュメントの冒頭から、埋め込みもうとする秘匿情報のビットに対応する文字数になる位置で改行コードを挿入していく。抽出時には、各行の文字数をカウントし、同じ対応表を用いて秘匿情報を抽出する。つまりこの方式では1行につき1ビットの情報を埋め込むことになる。図-2に、カバーテキストとステゴテキストの例を示す。1行当たりの行幅（各文字の字幅の合計）の変動を小さくしてドキュメントの見た目の不自



1 行文字数と埋め込み情報との対応表

図-1 改行位置の制御による情報ハイディング (処理の流れ)

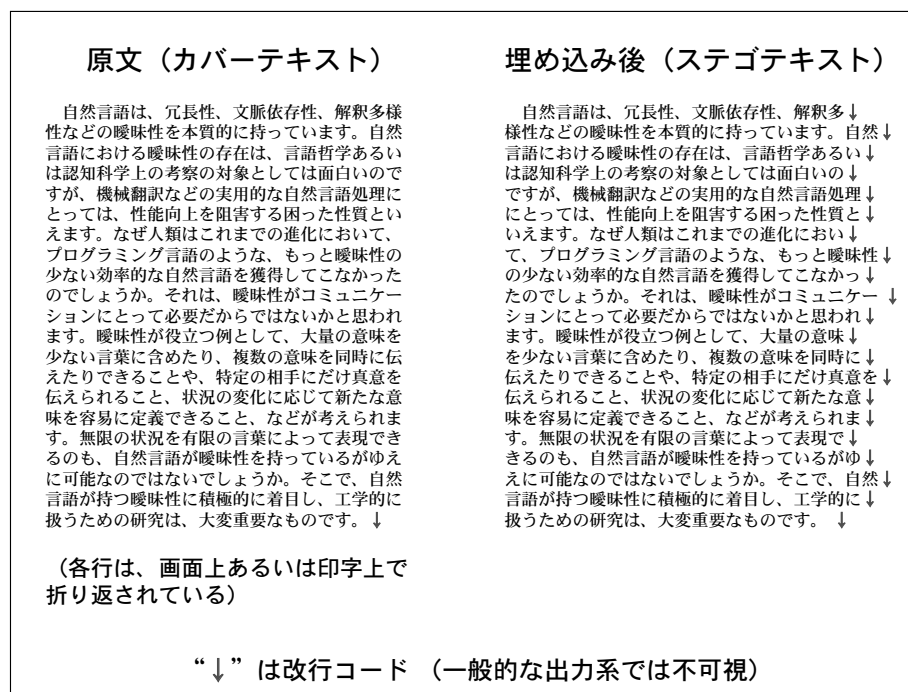


図-2 改行位置の制御による情報ハイディング (カバーテキストとステゴテキストの例)

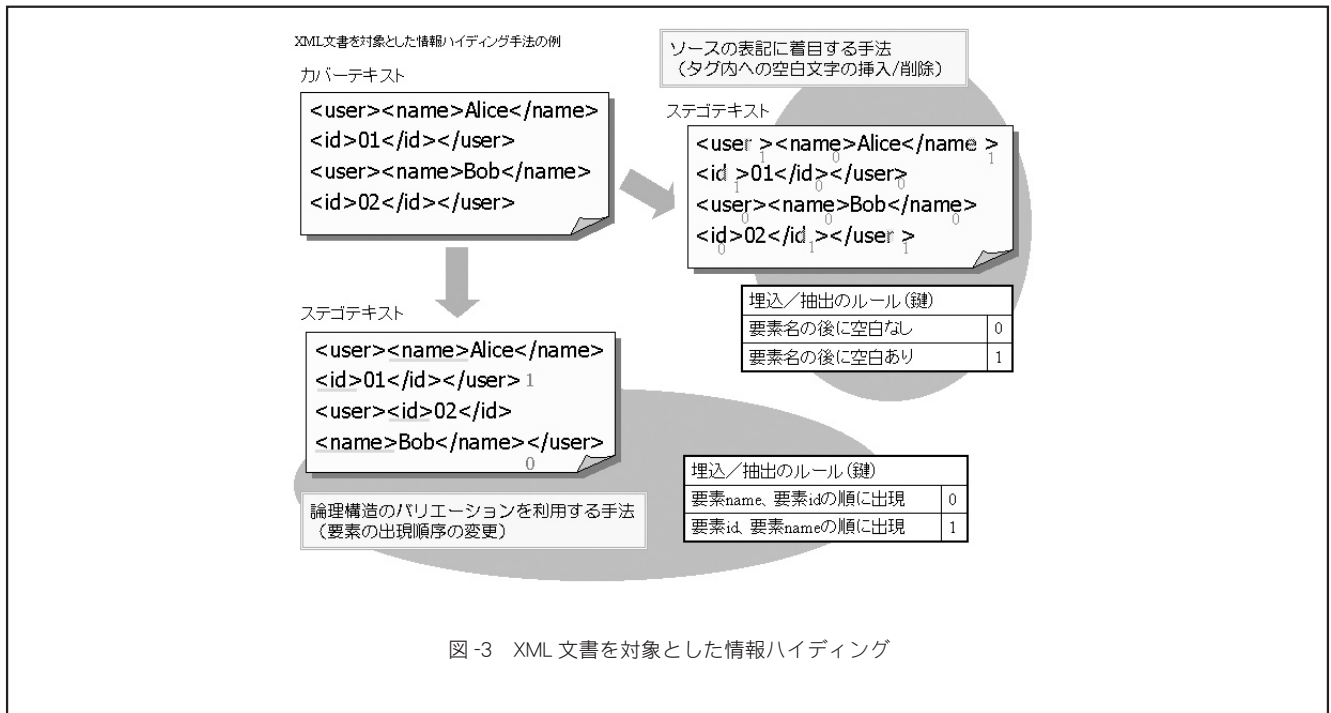
然さを少なくする戦略と、膠着言語とはいえず単語の途中などの不自然な位置での改行をなるべく回避する戦略とのトレードオフを考え、最も不自然さが少ない方法を選ぶ。この方式は、文面にまったく影響を及ぼさないため、改変困難な著作物をカバーテキストにする場合にも適用できる。またこの方式は、プレーンテキストへの文字レベルでの作為であると同時に、改行位置というレイアウト上の作為にもなっている。

なお、元となるカバーテキストが存在せず、埋め込みデータに従ってステゴテキストを無から生成する方式も、このタイプに属する。そのような方式として、uencode ファイルや PGP メッセージを、あたかも詩の

ような英文に変換する“Texto”や、バイナリデータを指定された文体の英文に変換する“NICETEXT”などのツールが提案されている⁶⁾。

■作為がハードコピー上に残らない情報ハイディング

ハードコピー上に残らない方式は、見た目に作為がまったく識別できないため見破られにくいものの、電子データから表示メディア（紙、画面など）に変換された時点で秘密情報が消去されるので、秘密情報の抽出時まで電子データのままで扱う利用法が前提となる。



この方式としては、英文をカバートテキストとし、複数の空白文字を各行末に挿入することにより情報を埋め込む“SNOW”と呼ばれる手法⁶⁾などが提案されている。SNOWでは、秘匿情報はハフマン符号化により圧縮して暗号化した後、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。そのほか、ヌルキャラクタをモルス信号に則ってテキストデータの中に配するFFEncodeというツールも提案されている⁶⁾。さらに、英文のLaTeX文書をカバートテキストとし、ソース中の本文の各行の単語の個数を加減することにより、情報を埋め込む手法も提案されている⁷⁾。

XMLなどの構造化文書への埋め込みも、基本的に作為がハードコピー上に残らない方式に属する。構造化文書を対象とする方式については、次の章で詳述する。

■構造化文書への情報ハイディング

ネットワーク上を流通するテキストの主要な形式としては、MS Word形式やAdobe PDF形式等のアプリケーション固有の書式で作成された文書や、HTMLやXML等のマークアップ言語で書かれた整形済みデータがあげられる。これらの文書はその内部に、文書の内容そのものの(本文)以外に、文書の論理的な構造に関する情報や体裁に関する情報を表現する部分を持っている。本章では、このような整形された文書を構造化文書として広く捉え、テキスト本文ではなく、文書構造を表す部分に対して適用可能な情報ハイディング手法について、大きく2種に分けて概説する⁸⁾。例としてXML文書を対象とした情報ハイディング手法を図-3に示す。

これらの手法の適用は出力結果(表示画面・印刷物)

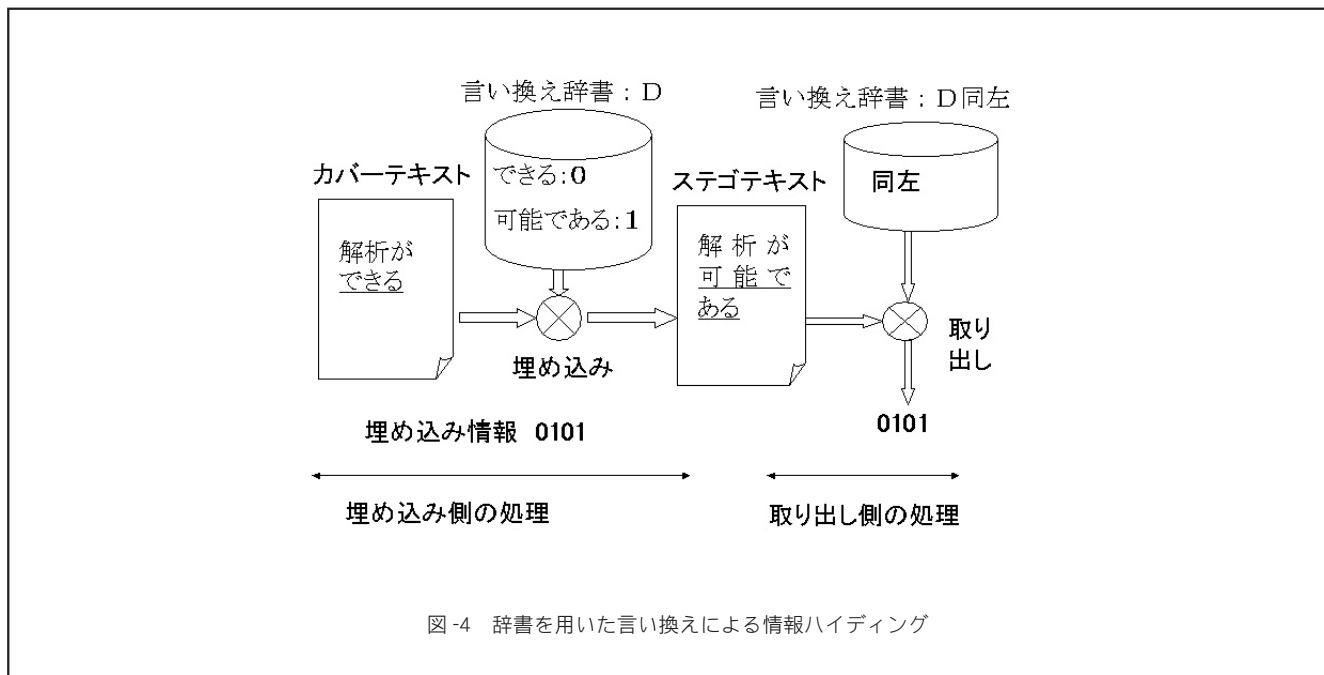
のみからは判断できない。広く使われているMS Word文書、Adobe PDF文書、HTMLについては、表示・印刷イメージとほぼ同じ形式で目視して編集や閲覧が行われる。また、データ交換に広く使われるXML文書についても、内容を確認する際には、円滑な参照や入力を助けるためのアプリケーションが利用されており、文書ソースを表示する機会はあまりない。構造化文書に対する情報ハイディング手法は、その適用がプログラム処理の奥に隠されてしまうため通常人目には気づかれにくいとも言える。

◇ソーステキストの表記に着目する手法

構造化文書においては、同じ内容・体裁・構造を示すための複数の表記がルール上許されていることも多い。カバートテキストを本来利用するための処理において、これらの表記のバリエーションに対し同一表示や同一印刷物が得られる場合には、バリエーションをステゴテキストとして埋め込む情報に対応付けた情報ハイディングが可能である。既存の研究事例としては、前述のSNOWなどがある。

文書への埋め込みの有無の検出は、文書構造ではなくソースの表記そのものを解析することで可能となる。同一のカバートテキストから生成された複数のステゴテキストを、文書構造の面で比較しても両者はまったく同一であると判断される。この手法で埋め込まれた情報は、文書本来の用途と無関係な部分の冗長性を利用して埋め込まれているため、文書の正規化や最適化等の処理により、ステゴテキストから失われてしまう可能性が高い。

XML文書においても、マークアップ表記上の揺らぎがあり、これを利用すれば、XMLに特化した情報ハイ



ディング手法を構成できる (図-3 右部)。要素や属性の冗長な表現、タグ内の空白文字の有無などは多くのアプリケーションにおいて処理上はバリエーションが同一視されるため、情報の埋め込みに利用できる。

◇論理構造のバリエーションを利用する手法

構造化文書を利用するシステムにおいて、異なる論理構造を持つ文書についても、同じ処理の結果 (たとえば同一の画面表示) が得られる場合がある。これら論理構造のバリエーションをステゴテキストとし、埋め込む情報に対応付ければ情報ハイディング手法を構成できる。

たとえば、文書中の複数の要素に対し、利用時にソート結果を表示する場合が考えられる。文書ソース上の要素の出現順序を変更すれば文書構造も変更されるが、表示上はこの変更は現れない。XML 文書においても、データ交換用途で利用される場面などでこの手法を適用できる (図-3 左下部)。

この手法ではカバーテキストの論理構造を埋め込む情報に対応させて変更したステゴテキストを生成する。文書の正規化や最適化により埋め込まれた情報が失われることはない。

■自然言語処理を応用した情報ハイディング

◇枠組みと分類

ここ 10 年間に自然言語処理の技術は大きく進歩し、日本語においては文を単語に分割し、英語においては変化形 (名詞の複数、動詞の人称変化、過去、過去分詞) から原形を求める形態素解析システムは高性能のものを容易に使えるようになってきた。辞書も大規模化、精密

化し、構文解析や自動言い換えの研究も進んでいる。これらの資源を利用する情報ハイディングが実現されている。現在、考えられている一般的な枠組みは図-4 のような構造である。言い換え辞書 (言い換えする言語表現の対 a, b に対して、情報 0,1 を割り当てた (a: 0, b: 1) の集合) をあらかじめ用意しておき、埋め込み側と取り出し (抽出) 側で共有する。埋め込み側では、カバーテキストを走査して、言い換え辞書に登録された表現を発見すると、その表現を使うか、それを言い換えた表現を使うかによって、埋め込み情報を埋め込んでステゴテキストを作る。図-4 の例では、カバーテキスト中の「できる」という表現に対して、埋め込みたい情報が 1 なので、言い換え辞書: D を見て、これを「可能である」という表現に言い換えてステゴテキストを作っている。取り出し側では、埋め込み側と同じ言い換え辞書: D を用いてステゴテキストから埋め込み情報を取り出す。

このシステムを方法論的に分類すると、(1) 生成型、(2) 言い換え型、と分かれ、言い換え型はさらに、意味を保存しない型、意味を保存する型に分かれる。一方、使用する技術の点からみると、(1) 事例ベース辞書つまり上記のような言い換え辞書による方法、(2) 構文的言い換えによる方法がある。以上の各分類について説明する。

◇生成型

無意味文だが、一見文らしい文字列を生成する方法であり、以下のシステムがある。

(1) 文→主語、述語 のような多数の書き換え規則の各々に 0 か 1 を割り当てておき、埋め込み情報に対応する規則を順次適用して文を生成するシステム。できた文は無意味ながら統語的には正しい。適用可能な書き換

え規則に曖昧さがあると、正しく埋め込み情報を取り出す方法に工夫が必要になる。

(2) male female male というタイプと、male= Tom 1, John 0, female= Alice 1, Betty 0 という辞書を用い、010 という埋め込み情報から、John Alice Tom を生成するシステム⁹⁾。タイプ1つにつき1ビット埋め込めるので多量の情報を埋め込める。

◇意味非保存の言い換え型

実際の自然言語文書をカバーテキストとして用い、言い換え辞書を適用する。ただし、言い換え辞書は意味を保存する言い換えに限らない。たとえば「情報 $\leftrightarrow$ データ」とか「研究 $\leftrightarrow$ 解析」程度の言い換えを行う。ステゴテキストが一瞥してテキストのように見えるならば文章の内容は問題ではないので、言い換え辞書の作成は容易である。ただし、ステゴテキストが日本語らしいものにするとなると、埋め込める場所は限定される。それを無視して、たとえば、「する $\leftrightarrow$ した」の言い換え、さらには能動態と受動態の言い換えなどの構文の変換などをすると、日本語としての質は明らかに低下する。さらに工夫する場合は、ある分野のエキスパートでない人には区別が付きにくい単語で置き換えることも考えられる。たとえば、セキュリティの専門家でなければ、「安全性 $\leftrightarrow$ 完全性」という言い換えをしても気づきにくい。

言い換え辞書を使わず、カバーテキストをSD式と呼ばれる意味表現形式に変換し、SD式が持つ意味的情報量を増減することによって情報を埋め込む手法¹⁰⁾も、この型に分類することができる。

◇意味保存の言い換え型

自然言語文のカバーテキストに意味を変えない言い換えを施して情報を埋め込む。意味を保存する場合、現在の自然言語処理技術では構文の変換は困難である。なぜなら、意味を保存するには係り受けなどの構文構造を完璧に認識しなければならないが、現在の技術では100%の認識精度には程遠いからである。実際、100%の認識には、文脈情報や常識知識まで必要になることが多く、実質的に不可能である。そこで、個別表現の意味が文脈に依存せず意味保存して変換できることを保証できる単語や常套句の言い換えの利用に限定されることになる。つまり、局所的な言い換えが実用的ということである。意味保存の言い換えであるから、日本語としての自然さは当然のこと、内容が正確に把握できることと、情報が埋め込まれていることが読み手に気づかれないことも求められる。中川らの実験¹¹⁾では、「できる $\leftrightarrow$ 可能である」「始める $\leftrightarrow$ 開始する」のような一般的な言い換えを用いた場合は日本語としての質が維持され、同一内容として理解されることが確認された。一方、「OS $\leftrightarrow$

オペレーティングシステム」のような専門用語の言い換えをすると、意味保存性と日本語の質の両者が劣化した。また、埋め込み箇所もテキスト1KBあたり数バイトと少ない。しかし、ソフトウェアの付属文書（マニュアル、使用許諾）においては著作権情報を埋め込むには十分であった。

■今後の展望

ドキュメントへの情報ハイディングを実現するための周辺技術の進歩には、目覚ましいものがある。まず、高解像度のレーザービームプリンタやディスプレイの普及により、レイアウトへの情報ハイディングが現実的に可能になってきた。そして、タグを自由に設計できるXMLの普及により、構造化文書への情報ハイディングが可能になった。さらに、形態素解析が一般化して単語単位での処理が容易になったことから、自然言語処理を応用した情報ハイディングが可能になった。特に自然言語処理については近年、意味を変えずに自動的にテキストの一部を言い換える技術の研究が盛んになっており、情報ハイディングへ容易に適用できるようになりつつある。

マルチメディア化が進んでいる現代においても、電子メールなどテキストによる情報交換はいまだ主流の位置を占めており、情報伝達手段としてのドキュメントの重要性は今後も変わらないと考えられる。したがってドキュメントへの情報ハイディングには、多くの応用が期待できよう。

参考文献

- 1) 滝澤 修, 山村明弘: 自然言語文を用いた秘密分散の提案, 情報処理学会コンピュータセキュリティシンポジウム (CSS2001), pp.343-348 (2001).
- 2) 松井甲子雄: 電子透かし的基础, 森北出版 (1998).
- 3) Anderson, R. J. and Petitcolas, F. A. P.: Information Hiding -An Annotated Bibliography, http://www.cl.cam.ac.uk/~fapp2/steganography/bibliography/Annotated_Bibliography.pdf (1999).
- 4) 瀬川典久, 村山優子, 宮崎正俊: 手書き入力装置の特性を利用した手書きステガノグラフィの提案, 情報処理学会 コンピュータセキュリティシンポジウム (CSS2002), pp.215-219 (2002).
- 5) 滝澤 修, 牧野京子: 改行位置の調整によるドキュメントへの情報ハイディングツールの開発, 情報処理学会 コンピュータセキュリティシンポジウム (CSS2002), pp.209-214 (2002).
- 6) 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> (1998).
- 7) 松本 勉, 糸山大志: 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).
- 8) 井上信吾, 村瀬一郎, 滝澤 修, 松本 勉, 中川裕志: XMLにおけるステガノグラフィ手法の提案, 電子情報通信学会 暗号と情報セキュリティシンポジウム (SCIS2002), pp.301-306 (2002).
- 9) Chapman, M. and Davida, G.: Hiding the Hidden: A Software System for Concealing Ciphertext as Innocuous Text, IJCS '97, pp.335-343 (1997).
- 10) 新見道治, 峯脇さやか, 野田秀樹, 河口英二: SD 式意味モデルに基づく自然言語文への情報ハイディング, 電子情報通信学会 暗号と情報セキュリティシンポジウム (SCIS2003), pp.965-970 (2003).
- 11) 中川裕志, 三瓶光司, 松本 勉, 柏木健志, 川口修司, 牧野京子, 村瀬一郎: 意味保存型の情報ハイディングー日本語文書への応用ー, 情報処理学会論文誌, Vol.42, No.9, pp.2339-2350 (2001).

(平成 15 年 2 月 10 日受付)