テキストへの情報ハイディング

中川 裕志* 小俣 祐介* 松本 勉†

1 はじめに

情報の内容を秘匿する暗号に対して情報の存在自体を秘匿する情報ハイディングの研究がさかんになってきている．しかし，これまでの情報ハイディングは画像を対象とするものが大部分であり，またテキストを対象にする場合でも，空白の位置を微妙にずらして情報をハイディングするなど，実質的には画像として扱っていた [1].

そこで，本論文では自然言語処理技術を利用したテキストベースの情報ハイディング方式の提案と開発したシステムの評価を報告する．具体的には，隠蔽可能な情報量とハイディングによって生成されたテキストの自然さの関係を実験的検討の結果から述べる．

2 テキストへの情報ハイディングシステム

2.1 一般的な枠組み

情報ハイディングとは，情報の存在自体を秘匿する技術であり，一般に二者によるコミュニケーションを前提とした技術である．図1に示すように，情報の埋め込み (embedding)・伝送 (transmitting)・抽出 (extracting) によって構成される．また，図1において，embedded data とは秘匿すべき情報を示し，cover data とは embedded data を埋め込む対象を示す．stego data とは，embedding によって embedded data が埋め込まれた cover data である．

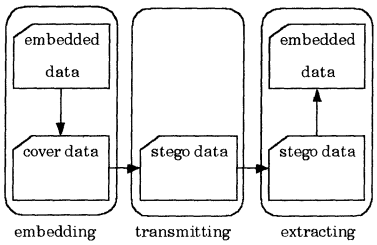

図 1: 情報ハイディングの枠組み

本研究では，cover data をテキスト (これ以降 cover text と呼ぶ) としたときの embedding や extracting の方式について提案し，embedding によって生成された stego data であるテキスト (これ以降 stego text と呼ぶ) に対する感知されにくさを評価する．

* 横浜国立大学工学部電子情報工学科
† 横浜国立大学大学院工学研究科人工環境システム学専攻

2.2 システム概要

テキストへの情報ハイディングシステムの方式や特徴を以下に述べる．

- 秘匿情報 (embedded data)e は図2のようにバイナリストリングとして入力され，送信者が指定した cover text C 中に埋め込まれて自然言語文 (stego text)S となって出力される．
- 秘匿情報は文章中の単語の置き換えによって行われる．そこで，あらかじめ別の言語リソースであるテキスト群から置き換えの対象となる単語を取り出し，各語に embedding する情報をビット単位で割り当てた辞書 D を用意する．
- この辞書を利用することで図2のようにテキスト C 中から変換の対象となる単語を取り出し，秘匿情報を持つような単語に置き換えることによって情報を埋め込む．

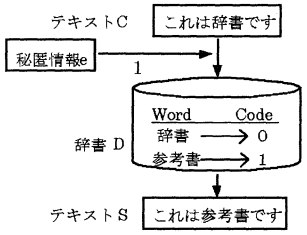

図 2: 辞書変換法による秘匿情報の埋め込み

- 埋め込まれた情報を抽出するには，図2の逆操作によって行われる．例えば，図2の例ならテキスト S 中の「参考書」という単語を取り出し，辞書 D の Word と Code の対応から秘匿情報 1 を extract する．

2.3 変換する言語要素の選択

辞書変換法によるテキストへの秘匿情報の埋め込みは，文章中の単語を変換することによって実現する．情報ハイディングが行われていること自体を検出しにくくするのが情報ハイディングシステムの要件なので，置き換えた語が文章中で不自然にならないことが肝要である．そのために，どの品詞が変換の対象として適切であるかを検討する．

各品詞の特徴を表1にまとめ，同じ品詞の他の単語で置き換えた場合を想定する．動詞・形容詞・副詞では語尾が活用するのでこれらを変換の対象とすると，

stego text の文法的不自然さをなくすためには，活用形までも考慮しなければならない．これは複雑な自然言語処理を必要としてしまうので可能ではあるが得策ではない．また，助詞を変換の対象とすると日本語としての文法的性質が乱され，テキストが扱っている分野の素人でも容易に stego text の不自然さを発見できるであろう．

次に，より多くの秘匿情報を埋め込むという点からは出現頻度が多いことと，1語あたりに埋め込むことのできる情報量を大きくするには語の種類数が多いことが望ましい．従って，文章構成に影響を及ぼさず，比較的出現頻度や種類数の多い名詞が，秘匿情報を埋め込む対象として最適であると考えられる．

しかし，名詞を無作為に置き換えただけでは置き換え後の単語が不自然になることがある．そこで我々は，できるだけ文の自然さを損なわないような秘匿情報の埋め込みによる変換を行うことを目的として 1) 複合名詞を利用する方法，2) 単名詞であっても置き換えが不自然さを引き起こしにくいような語に制限する方法を検討した．

表 1: テキスト中における各品詞の相対的な特徴

特徴＼品詞	名詞	動詞	形容詞 副詞	助詞
出現頻度	多い	少ない	少ない	多い
種類数	多い	多い	少ない	少ない
語尾の活用	無し	有り	有り	無し
変換後の 文法の乱れ	小さい	大きい	大きい	大きい

3 複合名詞の連接構造を保存した情報ハイディング法

3.1 複合名詞のパターン辞書構造

複合名詞とは，単名詞の連続である部分と複数の単名詞または複合名詞の間に「の」という接続助詞を含めた一連の語を指すことにする．日本語における複合名詞はそれを構成する末尾の単名詞が主辞という重要な性質を持っており [2]，その前方に連接する単名詞は末尾の名詞を修飾する形で存在する．よって，末尾の名詞を固定してその前方に連接する単名詞を他の名詞に置き換えることができれば，複合名詞全体の文法的性質は大きく変化しないと考えられる．

秘匿情報の埋め込みに用いる辞書を構築する手順として，文書集合から複合名詞の成分になっている単名詞を全て拾い出す．次に末尾の名詞を基準としてその前方に連接する単名詞のリストを構築する．さらにリスト中の各単名詞に対し同様の作業を繰り返すと，図3のように末尾の名詞を根とした木構造の名詞の連接関係が構築される．但し，A ～ J は複合名詞を構成する単名詞を表す．

このようにして全ての名詞の連接関係が得られた後，各単名詞の前方連接名詞に対して名詞数に従って図3のようにビットを割り当てる．

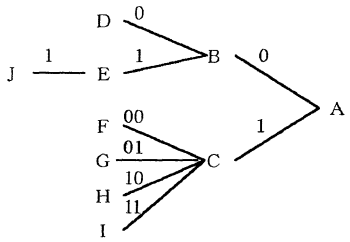

図 3: 末尾名詞を基準とした複合名詞パターン辞書

この辞書を用いることにより，テキストから複合名詞を抽出し，これを秘匿情報に従ったビット情報を持つ複合名詞に置き換えることにより秘匿情報の埋め込みが行われる．

3.2 複合名詞の抽出

テキストから特定の品詞を取り出すには，品詞情報を得ることができる形態素解析システムがよく利用される．しかし形態素解析は非常に重い処理であり，文章構成によっては誤解析が起こり，秘匿情報の埋め込まれた複合名詞を正確に抽出できるという保証がない．そこで，embedding や extracting で秘匿情報の埋め込まれた複合名詞を確実に抽出するために形態素解析システムを使わずに以下の方法を用いる．

(A) テキスト中から最も文末に出現する 3.1 の辞書に登録された複合名詞の末尾語を抽出する
(B) 末尾語に対する前方連接名詞のリストを辞書から抽出する
(C) 前方連接名詞が末尾語の前方に出現するかどうかリストを用いて調べる
(D) もし前方連接名詞が出現したらその名詞に対し (B) ~ (D) の作業を繰り返す
(E) もし前方連接名詞が存在しなければそこまで出現した名詞群を複合名詞とする

末尾語を抽出するには予め辞書中に複合名詞の末尾に出現した名詞のグループを用意しておく．また，辞書中の名詞には語の一部を包含するような名詞が存在することがある[1]．このような名詞が同一グループに存在した場合，文字列の長い方を優先して選択し，前方連接名詞の検索を行う．

この手法により，抽出された複合名詞が embedding によってどのような語に置き換わったとしても，複合名詞中の各名詞間には連接関係があるので辞書を利用すれば extracting で正確に秘匿情報を取り出すことができる．

3.3 embedding に関する評価結果

実験では，同一分野の技術論文から5つのテキストを cover text として取り出し，1つのテキストの

[1] ex.「数式」と「式」

みで作成した辞書Aと5つのテキストから作成した辞書Bによって各テキストのembeddingを行った．そのとき埋め込まれた秘匿情報量の統計を各cover textの平均で比較して，どの程度の量の秘匿情報が埋め込めたのかを検討する．

表2より秘匿情報をより多く埋め込むためにはcover textだけではなく，複数のテキストから辞書を作成した方がよいと思われる．その理由は，③の項目より置き換えられた複合名詞の構成する単名詞数がcover textのみの辞書のものと比較してほどんど変化がない．つまり，②,④の項目より置き換えることのできる複合名詞が増加し，1単名詞あたりの秘匿情報量が増えたため，全体の埋め込まれた情報量が増加すると考えられる．

表2: embeddingによる複合名詞の統計情報

利用した辞書	①	②	③	④	⑤
cover textのみで作成した辞書A	60.2	244 {467}	2.16	1.56	2.55
5つのテキストから作成した辞書B	95.7	314 {467}	2.19	1.99	4.10

① 埋め込まれた秘匿情報量 (byte)
② 置き換えられた複合名詞数 { 全複合名詞数 }
③ 複合名詞を構成する単名詞数
④ 複合名詞中の1単名詞あたりの秘匿情報量 (bit)
⑤ cover text 1KBあたりの秘匿情報量 (byte)

また，embeddingによって生成されたstego textの例を図4に示した．図では，秘匿情報の埋め込みによって置き換えられた複合名詞を [置き換えられた複合名詞／埋め込まれた秘匿情報] で示している．これより生成されたテキストは，一見した限りでは秘匿情報が埋め込まれていることが判らないと思われる．

[暗号技術 /01] を利用した [サービス提供 /0] 者は，登録しているユーザにだけサービスを提供したい場合がある．しかし，正当なユーザになりすまし不当にサービスを受けようとする攻撃者が存在する．
この [ときサービス提供 /00] 者 (認証者) は，サービスを要求したものが正当なユーザ (証明者) であるかの判定をする [第三者の署名方式 /11100] が必要となる．その [識別方式 /000] は，攻撃者によって容易に破られるものであってはならない．人間 - 機械間の [認証方式 /010] において，現在最も広く使われている方式はパスワード方式である．

図4: 辞書Bのembeddingによるstego text[3]

しかし，cover text 1KBあたりの埋め込むことのできる秘匿情報が非常に少ないことが欠点である．辞書作成に使用するテキストの数を増やしたり，cover textのサイズを大きくすることによってある程度の埋め込む秘匿情報を増やすことは可能だが，辞書作成にはなるべく同分野のテキストでないと置き換えた後の複合名詞が文章の内容と一致せず，stego textの不自然さが増大するおそれがある．また表2より，embeddingによって全ての複合名詞に秘匿情報が埋め込まれていないので，その語にも秘匿情報を埋め込み，複合名詞を構成する1単名詞あたりの埋め込むことのできる情報量を増やすことができればより多くの秘匿情報を埋め込むことができる．

この考え方に沿った埋め込む秘匿情報を増加する方法を次節で提案する．

4 複合名詞の連接構造を保存しない情報ハイディング法

4.1 複合名詞のグループによる辞書構造

ここでは，複合名詞を構成する単名詞すべてを1つのグループとして作り，各単名詞にビット情報を割り当てた辞書を用いる．この構造を表したものが図5となる．但し，A ～ Mは複合名詞を構成する単名詞を表す．この辞書を用いることで，テキスト中の複合名詞からそれを構成する単語数だけ辞書から取り出すことにより，複合名詞の置き換えが行われる．

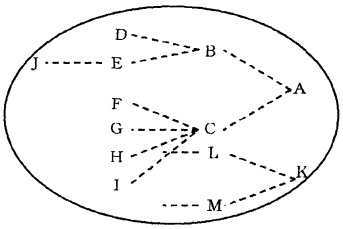

図5: 複合名詞のグループ辞書

4.2 embeddingに関する評価結果

3.3と同様の方法で評価を行った結果を以下に示す．表3より，連接情報を利用した辞書よりさらに多くの秘匿情報が埋め込まれていることが分かる．これは，1語当たりの秘匿情報量が大きいためと考えられる．置き換えられた複合名詞数が連接情報を用いた辞書よりも少なくなったのは，複合名詞の置き換えによって同じ単名詞が連続して出現してしまうような語を取り除いたためである．さらに，図6のように生成されるstego textは連接情報を利用した辞書と比較して不自然なものになってしまう．

表3: embeddingによる複合名詞の統計情報

利用した辞書	①	②	③	④	⑤
cover textのみで作成した辞書A	842	188 {467}	2.05	6.01	37.0
5つのテキストから作成した辞書B	1206	169 {467}	2.04	7.42	53.6

① 埋め込まれた秘匿情報量 (byte)
② 置き換えられた複合名詞数 { 全複合名詞数 }
③ 複合名詞を構成する単名詞数
④ 複合名詞中の1単名詞あたりの秘匿情報量 (bit)
⑤ cover text 1KBあたりの秘匿情報量 (byte)

まったく同じ入 [今後 /1010000][条件 /0110111] をもつモジュールを作ることは極めて困難であり，入出力の組が莫大な数になるため入出力の組すべてを記録することもほぼ不可能であるような性質を耐クローン性と呼ぶ．
本論文では人間の頭脳の耐クローン性を利用した対話型 [カウント /110101][契約 /0000000] 方式を検討する．ここで，あるモジュールが耐クローン性を持つとは，[頂点 /000000][対数 /0100000] と同等に働くものを作成するには，全ての入出力を取らなければならなく，また，そのようなテーブルを作るには莫大な時間を要することをいう

図 6: 辞書 B での embedding による stego text[3]

5 単名詞を利用する情報ハイディング法

5.1 単名詞の後方連接品詞別による辞書構造

複合名詞よりも出現頻度や種類数が非常に多い単名詞に秘匿情報を埋め込むことを考える．単名詞の場合，その名詞の文法的ないし意味的な性質によって後続する助詞が制限される．そこで，テキスト中の単名詞をその後に続く助詞や接尾辞などの品詞ごとにグループ化して，各グループごとにビット情報を割り当てる．この構造を表したものが図7となる．但し，A ～ M は単名詞を表す．次に cover text 中に単名詞が出現したら，その後の助詞と同グループ内の名詞に置き換えることによって，秘匿情報を埋め込む．しかし，複合名詞の連接関係や出現場所に対するグループとしての関係がないため，複合名詞の出現場所別による辞書よりも秘匿できる情報量は大きいが stego text はさらに不自然なものになることが予想される．

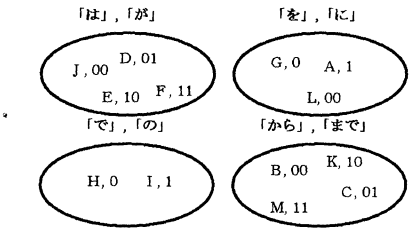

図 7: 単名詞の後方連接品詞別グループ辞書

5.2 embedding に関する評価結果

3.3 と同様の方法で評価を行った結果を以下に示す．表4より，3. の手法より多くの秘匿情報が埋め込まれているが，4.2 の結果よりは秘匿情報が埋め込まれていない．これは，図8を見ると分かるように秘匿情報を埋め込む対象が多く出現しているが，複合名詞の連接情報を持たない辞書ほど1グループ中の単名詞の数が多くないので，1語当たりの秘匿情報量が少なくなったためと考えられる．生成される stego text もまた，3. で述べた複合名詞の連接構造を保存する方式に比べて不自然なものになってしまう．また，辞書を作成するためのテキストの量を増やしても，埋め込むことのできる秘匿情報の増大には他の辞書よりも効果がなかった．これは，単名詞の種類数や置き換えられた単名詞の数にさほど変化がなかったためと考えられる．

表 4: embedding による単名詞の統計情報

利用した辞書	①	②	③	④
cover text のみで作成した辞書 A	548	814 {1035}	5.36	24.4
5 つのテキストから作成した辞書 B	694	851 {1035}	6.50	31.2

① 埋め込まれた秘匿情報量 (byte)
② 置き換えられた単名詞数 { 全単名詞数 }
③ 1 単名詞あたりの秘匿情報量 (bit)
④ cover text 1KB あたりの秘匿情報量 (byte)

情報通信技術を [記述 /0101011] した [レイアウト /11 001] の提供者は，[記載 /0101001] している [パラメータ /11100] にだけ [パスワード /11111] を [解析 /0010 010] したい [評価 /1101001] がある．しかし，正当な [リクエスト /00101] になりすまし不当に [コンピュータ /00101] を受けようとする攻撃者が [抽出 /0111] する．
このときサービス提供者 (認証者) は，[コンパイル /101001] を [統一 /0110100] したものが正当な [テーブル /01110](証明者) であるかの [処理 /10010111] をするための認証方式が必要となる．その認証方式は，攻撃者によって容易に破られるものであってはならない．[入手 /11000111]- 機械間の認証方式において，現在最も広く使われている [変更 /010010] はパスワード方式である．

図 8: 辞書 B での embedding による stego text[3]

6 まとめ

提案した3通りの方法の評価結果から埋め込むことのできる秘匿情報が増えるほど生成されるテキストの不自然さが増加することがわかる．従って本研究では，埋め込むことのできる秘匿情報量を減らさず，如何に自然なテキストを生成するかが今後の課題となる．

謝辞

本研究は情報処理振興事業協会のプロジェクトの一環として行われました．

参考文献

[1] 株式会社三菱総合研究所．「インフォメーションハイディングの技術調査」報告書．Technical report, 1998 年 2 月.

[2] Takao Gunji. Japanese phrase structure grammer. *Reidel Dordrecht*, 1987.

[3] 林修一．耐クローン性に基づく対話型個人識別方式の実装．横浜国立大学工学部電子情報工学科卒業論文, 1997.