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

中川裕志[†] 木村浩康[††]
三瓶光司[††] 松本勉[†††]

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．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．この研究では従来とはまったく異なり，テキストの内容自体を書き換えることによって情報ハイディングを行う技術を提案する．当然，ハイディングによって埋め込む情報量が増えるにつれてテキストは不自然になる．このことを我々が実装した情報ハイディングシステムによって評価した結果などについても報告する．

Information Hiding for Japanese Text Based on Replacing Words with Dictionary

Hiroshi Nakagawa,[†] Hiroyasu Kimura,[††] Koji Sampei[††]
and Tsutomu Matsumoto[†††]

Information hiding becomes increasingly focused on in the area of information security. We propose technique which hides information into text by replacing words using a dictionary being consists of pairs of words in this paper. The first of the pair is a word to be replaced, and the second of the pair is the word with which the first word replaces. We propose three kinds of dictionaries based on Japanese word formation. We also evaluate the text in which information is hidden by the amount of hidden information and also unnaturalness of the text by hand.

1. はじめに

近年，インターネットの普及とともにインターネットにおけるセキュリティ確保の必要性が高まっている．そこで，情報の存在を隠蔽する技術である情報ハイディングが注目されている．情報ハイディング技術の目的は，情報の存在自体を隠蔽することであり，第三者に情報の存在を感知されないことを目的とする．これに対し暗号技術の目的は，情報の意味を隠蔽することを目的としており，情報ハイディング技術と暗号技術は，それぞれ情報の隠蔽する側面が異なっている．現在，情報ハイディングに関する研究は，画像分野における電子透かし技術の成熟が進んでいるが，テキストを対象にした情報ハイディングは事例が少なく，情報ハイディングの定義や秘匿可能な情報量に関する理論面においても研究の余地が残されている．

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への情報ハイディング方式に関する我々の提案と開発したシステムの評価を報告する．

† 東京大学情報基盤センター
Information Technology Center, The University of Tokyo
†† 横浜国立大学工学部電子情報工学科
Division of Electrical and Computer Engeneering, Yokohama National University
††† 横浜国立大学大学院人工環境システム学専攻
Division of Artificial Environment and Systems, Yokohama National University

2. 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 を埋め込む情報を示す．stego data とは，埋め込みによって embedded

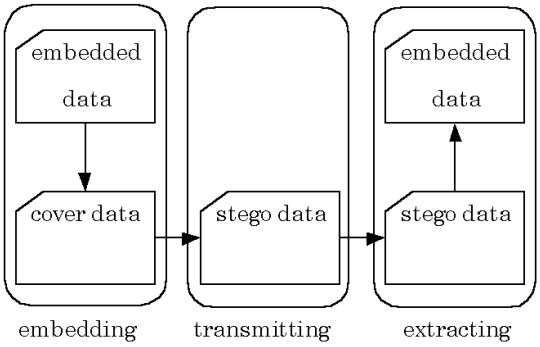

図 1 情報ハイディングの枠組み
Fig. 1 Scheme of information hiding.

data が埋め込まれた cover data である．

本研究では，cover data をテキスト（これ以降カバーテキストと呼ぶ）としたときの埋め込みや抽出の方式について提案し，情報の埋め込みによって生成された stego data であるテキスト（これ以降ステゴテキストと呼ぶ）に対する攻撃者からの感知されにくさを評価する．なお，ここでは，カバーテキストの意味内容が秘匿情報の埋め込みによって変化しないことを保証しない．

2.2 従来のテキストへの情報ハイディング法

秘匿情報を埋め込む対象がテキストである情報ハイディングは主に 3 つの方式がある．

(1) ホワイトスペース法[1),2)]
ページ設定上での文字間や行間の空白やスペースを制御することにより秘匿情報を埋め込む．

(2) 辞書変換法[3),4)]
あらかじめ用意された文法構造と各単語に秘匿するデータに対応する情報が割り当てられた辞書を用いることにより秘匿情報を埋め込み，自然言語風のステゴテキストを生成する．

(3) 文字埋め込み法[5),6)]
スペースやタブなどのヌルキャラクタを単語間または行末に埋め込むことにより，秘匿情報を埋め込む．

(1) の方式は本質的に画像への情報ハイディング方式であり，テキスト自体には情報が隠蔽されていない．また，(2)，(3) のどちらの方法においても生成されるステゴテキストが不自然なものになる．よって，何らかの情報が隠蔽されていることが単語や文字などの出現頻度情報を調べることによって検査する機械的な検出や人間が実際に読むことによる人為的な検出によって容易に発見されてしまうおそれがある．

そこで本研究では，辞書変換法を改善することにより人間が読んだ場合でもテキストの不自然さを検出されにくい，ないしは，検出に時間がかかるテキストへの情報ハイディング方式について検討する．

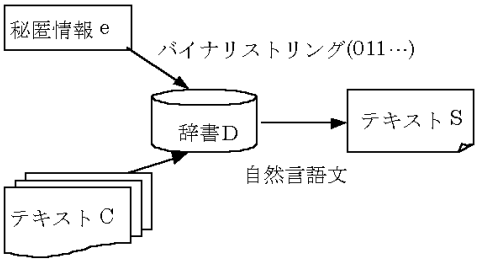

図 2 秘匿情報の埋め込み（embedding）の動作概要
Fig. 2 Embedding data into cover text.

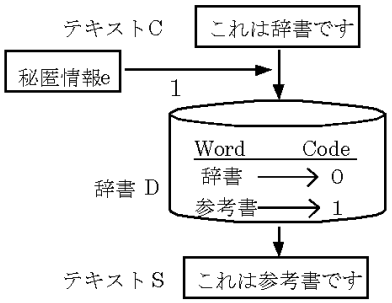

図 3 辞書変換法による秘匿情報の埋め込み
Fig. 3 Embedding data by replacing words with dictionary.

3. テキストへの情報ハイディングシステム

3.1 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 e は**図 2** のようにバイナリストリングとして入力され，カバーテキスト C 中に埋め込まれることにより自然言語文（ステゴテキスト）S を出力する．
- 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，各語にビット情報を割り当てた辞書 D を用意しておく．
- この辞書を利用することで**図 3** 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．
- 秘匿情報 e を埋め込む際に用いられる辞書を D，e が埋め込まれるカバーテキストを C とした場合，秘匿情報の埋め込みとは，D，C，e から S を作

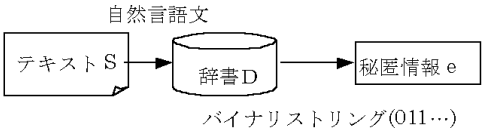

図4 秘匿情報の抽出（extracting）の動作概要
Fig. 4 Extraction embedded data from stego text.

り出すことである．

- 埋め込まれた情報を抽出するには，**図4**のように辞書Dのみが必要でありカバーテキストCは必要としない．

ここで述べる辞書変換法は図2の例からも分かるように，秘匿情報の埋め込みによって，単語が変化するため，元のカバーテキストCの意味は埋め込みによって変化してしまう．そこで，情報ハイディングシステムとしての評価はステゴテキストSとカバーテキストCの意味同一性ではなく，ステゴテキストSの自然さの程度である．

3.2 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．一般の文章は様々な品詞で構成されているので，どれを変換の対象とするかが問題となる．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．

一般の文章は主に名詞・動詞・形容詞・副詞・助詞などの品詞で構成されている．各品詞の特徴を**表1**にまとめた．一番下の欄は同じ品詞の他の単語で置き換えた場合の文法の乱れの大小を示す．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，ステゴテキスト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，素人でも容易にステゴテキスト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語あたりに埋め込め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．したがって，文章構成に影響を及ぼさず，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文章が意味的に不自然になる．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
Table 1 Linguistic nature of each POS in text.

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	なし	あり	あり	なし
変換後の 文法の乱れ	小さい	大きい	大きい	大きい

4. 複合名詞の連接構造を保存した情報ハイディング法

4.1 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[7]，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，まず対象とするカバーテキストと同分野の文書を複数集め，これらの文書から複合名詞の成分になっている単名詞，または単名詞の連続である複合名詞をすべて拾い出す．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の前方に連接する単名詞を取り出す．すると，**図5**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．ただし，A〜Jは複合名詞を構成する単名詞を表す．

このようにしてすべ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5のようにビットを割り当てる．ここで，名詞数がNのとき，割り当てられるビット数nは次のようになる．

$$n = \lceil \log_2 N \rceil$$

またこの辞書を用いることにより，次の手順で秘匿情報をカバーテキストに埋め込む．

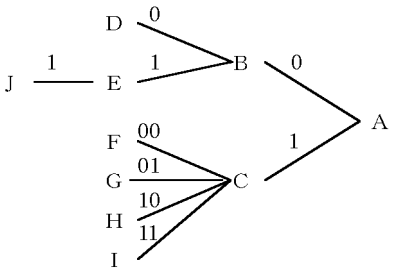

図 5 末尾名詞「A」を基準とした複合名詞パターン辞書
Fig. 5 A representation of compound noun ending the noun "A" in our dictionary.

(1) 4.2 節に述べる方法でカバーテキスト中から複合名詞を取り出す.
(2) 複合名詞中の末尾単名詞，あるいはすでに (1) で置き換えのすんでいる単名詞から前方連接する名詞の候補を辞書から取り出す.
(3) 前方連接名詞の候補から秘匿する情報を持つ名詞に置き換える.
(4) 置き換えた名詞に対して (2)〜(4) の作業を，元の複合名詞を構成する単名詞数になるまで繰り返す.

なお本方式では，カバーテキスト中で同一であった複合名詞がステゴテキスト中でも同一の複合名詞に置換させること，および，カバーテキストのサイズとステゴテキストのサイズが同一となることを保証しない．前者は，同一のものに統一してしまうと，秘匿可能な情報量が減少し，また，埋め込み後の複合名詞の不自然さはそれぞれ異なるため，不用意に統一してしまうと，明らかに不自然な語が複数回出現するといったことが考えられるからである．後者については，日本語において名詞，特に漢字で構成される名詞は，2 文字のものが大部分であり，カバーテキストとステゴテキストのサイズが大きく異なることはないためである☆.

4.2 複合名詞の抽出

複合名詞のパターン辞書を構築する際，複合名詞の抽出はテキストを文章を構成する品詞単位に分割し，その品詞情報を得ることのできる形態素解析システムが利用される．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，埋め込みや抽出では秘匿情報の埋め込まれた複合名詞を確実に抽出する以下の方法を用いる.

埋め込みや抽出で用いられる辞書は，カバーテキストを含む複合名詞中の名詞の連接情報が収められている．よって，テキスト中の名詞から辞書を利用して連接情報を調べれば複合名詞を抽出することが可能となる．その手順はテキスト中の各文に対して，以下の (1)〜(5) を適用するものである.

(1) 最も文末に出現する複合名詞の末尾語を辞書に登録された単語とのパターンマッチングによって，抽出する.
(2) 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.
(3) 前方連接名詞が末尾語の直前に出現するかどうかリストを用いて調べる.
(4) もし前方連接名詞が出現したらその名詞に対し (2)〜(4) の作業を繰り返す.
(5) 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が複合名詞となる．抽出した複合名詞の前方の部分に対して (1) からの作業を繰り返す.

末尾語を抽出するにはあらかじ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．また辞書の構成上，各名詞は前方連接の関係しか分からず，末尾だけではなく複合名詞の中間に存在する名詞もあるので，グループ中の名詞が文章に複数存在した場合，最も文末に出現した名詞を末尾語とする．この手法で複合名詞の抽出を行うと末尾のものから取り出されるので，埋め込みや抽出でも末尾の複合名詞から順に秘匿情報の埋め込みや抽出が行われる．また，辞書中の名詞にはたとえば「数式」と「式」のように語の一部を包含するような名詞が存在することがある．このような名詞が前方連接名詞や末尾語として存在した場合は，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.

この手法によって抽出された複合名詞が埋め込み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，辞書を利用すれば埋め込みに用いられたすべての複合名詞を含む集合を抽出できる．実際に埋め込みに用いられた複合名詞と，用いられなかった複合名詞が同一になる場合，前述の方法では区別が不可能であるが，埋め込みの段階で，実際の秘匿情報に始まりと終わりを示すビットパターンを付加することで，正確に秘匿情報を取り出すことができる.

4.3 埋め込みに関する評価結果

ここでは，5 つのカバーテキスト[8)〜12)]に対しカバー

☆ もっとも，ファイルサイズが変化すると考えられる単名詞を置き換える方式であっても 1 KB あたり 5〜6 byte 程度の変化にとどまった.

テキストのみから作成した辞書 A とカバーテキストを含む 5 つのテキストから作成した辞書 B で各カバーテキストの埋め込みを行った．そのとき埋め込まれた秘匿情報や複合名詞の統計を 5 つのカバーテキストの平均で比較して，どの程度の量の秘匿情報が埋め込めるのかを検討する．5 つのテキストの内容は同一分野の技術論文であり，テキストサイズは 15〜30 KB のものを利用した．

表 2 より秘匿情報をより多く埋め込むためにはカバーテキストだけではなく複数のテキストから辞書を作成した方が良いと思われる．その理由は，複数のテキストを辞書作成に利用することによって置き換えることのできる複合名詞が増加し，置き換えのパターンの増加によって複合名詞を構成する 1 単名詞あたりの秘匿情報量が増えるからである．

埋め込みによって生成されたステゴテキストの例を**図 6** に示した．この図では，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．この例では，一見した限りでは秘匿情報が埋め込まれていることが分からないと思われる．

表 2 4 章で提案した方法による埋め込み情報量
Table 2 Amount of embedded information by the method proposed in Section 4.

利用した辞書	1	2	3	4	5
カバーテキスト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) カバーテキスト 1 KB あたりの秘匿情報量（byte）

[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．単純な [認証方式/010] による [応用方式/111] は覗き見・盗聴といった攻撃に対して弱いという欠点がある．

図 6 4 章で提案した方法による文書 8) への埋め込みで生成したステゴテキストの一部
Fig. 6 A part of stego text created by embedding data into document 8) by the method proposed in Section 4.

カバーテキスト 1 KB あたりの埋め込むことのできる秘匿情報が 5 つのテキストから作成した辞書でも 4 byte と非常に少ないことが分かる．辞書作成に使用するテキストの数を増やしたり，カバーテキスト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，ステゴテキストの不自然さが増大するおそれがある．また表 2 より，埋め込みによってすべ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章で提案する．

5. 複合名詞の連接構造を保存しない情報ハイディング法

5.1 複合名詞の出現場所別による辞書構造

4 章で述べた方法では，複合名詞の連接関係を利用した辞書において，名詞の前方に連接する名詞の数に限りがあったので，複合名詞中の 1 単名詞あたりの情報量が少なくなってしまった．そこで，複合名詞を構成する各単名詞に対し，名詞が出現した場所別にそれぞれのグループを作り，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．ただし，A〜M は複合名詞を構成する単名詞を表す．

5.2 埋め込みに関する評価結果

4.3 節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，さらに 1 語あたりの秘匿情報量が大きいためと考えられる．しかし，**図 8** のように生成されるステゴテ

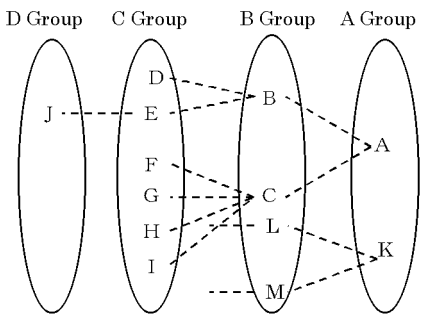

図 7 複合名詞の出現場所別グループ辞書
Fig. 7 Grouped nouns according to their position in the compound noun.

表 3 5 章で提案した方法による埋め込み情報量
Table 3 Amount of embedded information by the method proposed in Section 5.

利用した辞書	1	2	3	4	5
カバーテキスト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) カバーテキスト 1 KB あたりの秘匿情報量（byte）

[とおり発信入出力/0110110111000001111100010 1] を利用した [数の問い合わせ/111011001110001111] 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 [発展型情報/110010100010101100] 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 [相当部分何者/11001111000100010101110010] が必要となる．その [ブラックリスト言語/110001110000010101] 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 [観測の問い合わせ/111011001110001101] 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．単純な [大別一致/0000010010001011010] による [統一制能力/1111000100010010101000 1010] は覗き見・盗聴といった攻撃に弱いという欠点がある

図 8 5 章で提案した方法による文書 8) への埋め込みで生成したステゴテキストの一部
Fig. 8 A part of stego text created by embedding data into document 8) by the method proposed in Section 5.

キスト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．

6. 単名詞を利用する情報ハイディング法

6.1 単名詞の後方連接品詞別による辞書構造

単名詞に秘匿情報を埋め込むことを考える．しかし，普通に置き換えただけでは，ステゴテキストが非常に不自然になってしまうことは，先に述べたとおりである．単名詞の場合，その名詞の意味によって後続する助詞が制限されることがある．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 9** となる．ただし，A〜M は単名詞を表す．次にカバーテキスト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ステゴテ

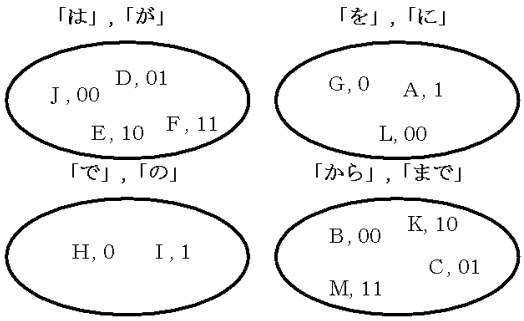

図 9 単名詞の後方連接品詞別グループ辞書
Fig. 9 Dictionary in which nouns are divided into some groups according to POS of next word.

表 4 6 章で提案した方法による埋め込み情報量
Table 4 Amount of embedded information by the method proposed in Section 6.

利用した辞書	1	2	3	4
カバーテキスト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) カバーテキスト 1 KB あたりの秘匿情報量（byte）

情報通信技術を [追加/11010000] した [変化/0110100] の [成功/0010100] 者は，[着目/0001001] している [規定/10001001] にだけ [滞在/0100001] を [工夫/0100101] したい [条件/011100] がある．しかし，正当な [安全性/0001100] になりすまし不当に [送受/0001010] を受けようとする [実用/11001110] 者が [リスト/0000101] する．
このときサービス提供者 ([紹介/11010101] 者) は，[周回/0000010] を [許可/10111101] したものが正当な [必要/10001111] ([改変/10110111] 者) であるかの [工夫/0100101] をするための認証方式が必要となる．その認証方式は，[抑止/0111110] 者によって容易に破られるものであってはならない．[関数/0001100]-[根本/10111110] 間の認証方式において，[コンピュータ/0010100] 最も広く使われている [インタフェイス/011101] はパスワード 方式である．単純なパスワード 方式による個人識別方式は覗き見・[サーバ/00110] といった [位置/10100110] に対して弱いという [人間/000001] がある．

図 10 6 章で提案した方法による文書 8) への埋め込みで生成したステゴテキストの一部
Fig. 10 A part of stego text created by embedding data into document 8) by the method proposed in Section 6.

キストはさらに不自然なものになることが予想される．

6.2 埋め込みに関する評価結果

4.3 節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，4 章や 5 章の手法よりさらに多くの秘匿情報が埋め込まれている．これは，**図 10** 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，さらに 1 語あたりの秘匿情報量が大きいためと考えられる．しかし，生成されるステゴテキストは前の 2 つの手法と比較して非常に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あまり効果がなかった．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．

7. ステゴテキストの主観評価実験

人間の主観による検証とは，人間が見てその文章が不自然と思うかどうかにすべて集約される．そこで，人間が不自然さに気づくまでの時間を計るという方法を用いて評価した．すなわち，人間の主観による検証の方法として，情報ハイディングされたテキストを，まったく情報ハイディングシステムの内容を知らない人に読んでもらって，情報ハイディングされていることに気がつくまでの時間を計った．

7.1 実験結果

カバーテキストとして論文と新聞[13]のテキストを用意した．論文，新聞の両リソースそれぞれ 10 種類のテキストを用いた．すべてのカバーテキスト全域に対し，情報を埋め込んだ．方式が 3 通り存在するので，全部で 30 種類の埋め込みをされたステゴテキストが揃う．これをそれぞれ何人かの人に読んでもらい，不自然さに気づくまでの時間を計る．今回の実験では情報系の学生 5 人に読んでもらった．被験者には不自然かどうかを判断するよう指示し，そのテキストに情報が埋め込まれているなど，本システムにかかわる情報はまったく与えなかった．また，ステゴテキストを査読する順番は被験者ごとにランダムに設定した．結果を**図 11** と**図 12** に示す．図 11 と図 12 はそれぞれ論文と新聞に対してのデータである．以下の図表においては，4 章，5 章，6 章で提案した各方式を，4 方式，5 方式，6 方式と記する．図 11 において 6 方式が一番多く情報が埋め込めこめることが分かるが，非常に不自然であることが分かる．4 方式と 5 方式はほとんど違いがなく，6 方式よりも埋め込める情報量は少ないが，不自然さが小さいということがいえる．図 12 においてもだいたい図 11 の場合と同じことがいえるが，図 11 では 4 方式と 5 方式に違いがあまりなかったのに対して，図 12 の場合では 5 方式の方が 4 方式よりも若干自然なテキストに置き換えられているということが分かる．

また，統計的な結果として**表 5** と**表 6** に示す．評価として不自然さに気づくまでの時間の分散，標準偏差，平均の 3 つのパラメータを求めた．平均において 6 方式を除いて不自然さに気づくまでの時間は論文の場合の方が長い．これは，新聞はある程度の一般常識があれば，不自然であることに気づきやすいということが考えられる．それに対し，論文の場合は専門知識

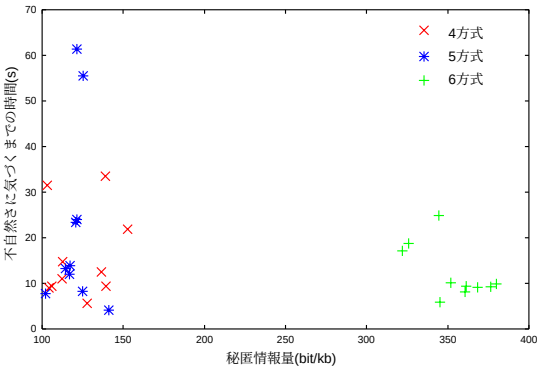

図 11 埋め込んだデータ量と検出に要した時間（論文の場合）
Fig. 11 Relation between amount of embedded data and required detecting time (Technicalpaper).

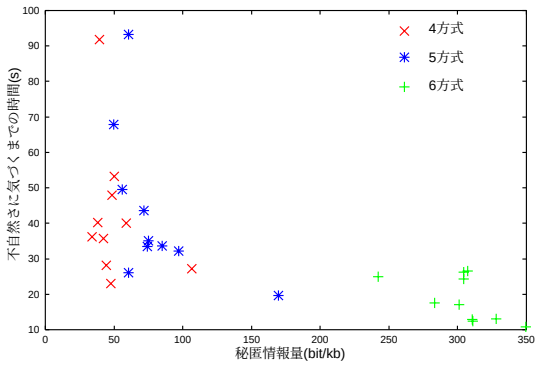

図 12 埋め込んだデータ量と検出に要した時間（新聞の場合）
Fig. 12 Relation between amount of embedded data and required detecting time (Newspaper).

表 5 不自然さを判定するまでの時間の統計（論文）
Table 5 Statistics of time to find unnaturalness (Technicalpaper).

	4 方式	5 方式	6 方式
標準偏差	28.57612	28.78816	9.777318
平均（sec）	42.3674	43.4361	18.5888

表 6 不自然さを判定するまでの時間の統計（新聞）
Table 6 Statistics of time to find unnaturalness (Newspaper).

	4 方式	5 方式	6 方式
標準偏差	18.63127	28.55533	28.55533
平均（sec）	15.8527	22.3471	22.3471

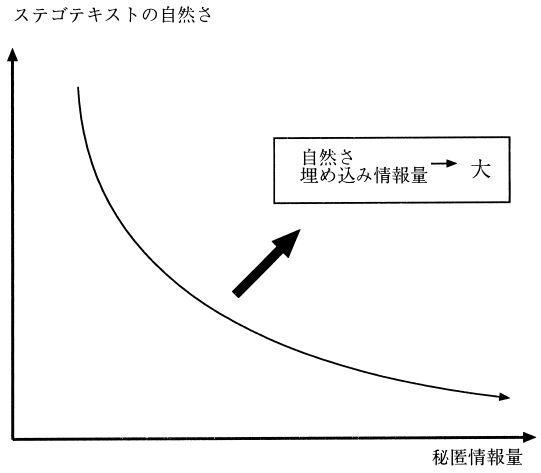

図 13 秘匿情報量とステゴテキストの自然さの関係
Fig. 13 Relation between amount of embedded data and naturalness of stego text.

が要求されるので不自然なテキストであっても不自然であることに気づきにくいということが考えられる．

これらの実験以外にまったく情報を埋め込んでいないテキストに対しても，不自然さに気づくまでの時間を計る実験と同様の実験も行った．論文の場合は 13 分，新聞の場合は 15 分で不自然かもしれないと誤認した被験者もいたが，他は同様の実験の結果，不自然ではないと判断した．よって，この実験では自然なテキストは自然と見なされていることも確認された．

8. 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分かる．これを，グラフに表すと**図 13** のような関係が得られることになる．よって本研究としては，図 13 の右上を目指す．すなわち，埋め込むことのできる秘匿情報量を減らさずにいかに自然なテキストを生成するかが今後の課題となる．さらに，計算機で機械的にステゴテキストの不自然さを判定する方法の開発も今後の課題である．

謝辞 本研究は情報処理振興事業協会（IPA）の情報セキュリティ関連事業（平成 10 年度）援助により行われました．本研究を進めるにあたって，ご助言いただいた三菱総合研究所の村瀬一郎氏に感謝いたします．また，本研究の評価実験のための文書リソースを提供していただいた松本研究室の糸山大志氏，池田竜郎氏，同じく評価実験に協力していただいた中川研究室の内間木新也氏，ならびに本研究の初期段階で尽力していただいた小俣祐介氏に感謝いたします．

参 考 文 献

1) Brassil, J., Low, S., Maxemchuk, N.F. and O'Gorman, L.: Hiding Infomation Documents Images, *Conference on Infomation Sciences and Systems* (*CISS-95*) (1995).
2) Brassil, J. and O'Gorman, L.: Watermarking Document Images with Bounding Box Expansion, *Info Hiding 96*, pp.227–235 (1996).
3) Chapman, M. and Davida, G.: Hiding the Hidden: A Software System for Concealing Ciphertext as Innocuous Text, *ICICS'97*, pp.335–345 (1997).
4) Wayner, P.: Mimic functions, *Cryptologia*, Vol.XVI, No.3, pp.193–214 (1992).
5) Maxemchuk, N.: Electronic Document Distribution, *AT&T Technical Journal*, Vol.73, No.5, pp.74–80 (1994).
6) Low, S., Maxemchuk, N., Brassil, J. and O'Gorman, L.: Document Marking and Identification using Both Line and Word Shifting, *Infocom 95* (1995).
7) 小俣祐介：テキストへの情報ハイディング方式に関する研究，横浜国立大学修士論文 (1999).
8) 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).
9) 加藤 功：質問の文字数を抑えた対話型個人識別方式の研究，横浜国立大学工学部電子情報工学科卒業論文 (1998).
10) 久保田浩美：耐クローン性に基づく認証方式の安全性評価，横浜国立大学卒業論文 (1997).
11) 水谷 亮：匿名ユーザの所属無効性を検出できる所属証明方式，横浜国立大学卒業論文 (1994).
12) 堤 暢彦：覗き見に強い対話型個人識別方式の操作性の改善に関する研究，横浜国立大学卒業論文 (1998).
13) 毎日新聞社：CD（毎日新聞'97 データ集），日外アソシエーツ (1998).

(平成 11 年 11 月 29 日受付)
(平成 12 年 6 月 1 日採録)

中川 裕志（正会員）

昭和 28 年生．昭和 50 年東京大学卒業．昭和 55 年同博士課程修了．工学博士．昭和 55 年より横浜国立大学勤務．平成 11 年 8 月より東京大学情報基盤センター教授．言語情報処理の研究に従事．

木村　浩康

昭和 48 年生．平成 10 年横浜国立大学卒業．平成 11 年同大学大学院工学研究科電子情報工学専攻修士課程在籍．

三瓶　光司

昭和 52 年生．平成 11 年横浜国立大学卒業．平成 11 年同大学大学院工学研究科電子情報工学専攻修士課程在籍．

松本　　勉

昭和 33 年生．横浜国立大学卒業．昭和 61 年東京大学博士課程修了．工学博士．昭和 61 年より横浜国立大学工学部電子情報工学科勤務，平成元年同助教授，平成 8 年 4 月より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助教授．情報セキュリティの研究に従事．