3-9　テキスト秘密分散法の研究

3-9 Secret Sharing Scheme Using Natural Language Text

滝澤　修　　山村明弘　　牧野京子
TAKIZAWA Osamu, YAMAMURA Akihiro, and MAKINO Kyoko

要旨

視覚復号型秘密分散法の考え方を自然言語テキストに適用した情報ハイディング手法を提案する。本手法は、日本語テキストを対象とし、複数枚の分散テキスト（share text）を重ね合わせて、上層から下層に読んでいくと、その文字列の中に秘密テキスト（secret text）が現れるようにするものである。分散テキストを重ね合わせることは簡単な機械処理によって実現できる。重ね合わせて得られた文字列の中から秘密テキストを抽出する際には、意味を持たないフレーズが一文字の形態素の連鎖になる割合が多い性質を利用して、形態素解析器を使用する。分散テキストは、テキストデータベースを用いて既存の文をつなげることによって生成され、一見自然なテキストに見せかけるための拡張性を有している。自然な分散テキストを生成するために、分散テキストを構成する文間の結束性の概念を用いる方法についても議論する。

Modifying the idea of the visual cryptography, we propose a method of sharing a secret key using natural language texts. Our target here is restricted to Japanese texts. Each participant obtains a share, which is a Japanese text in our scheme. When a certain number of participants retrieve the secret key, they supply their shares and pile up these natural language texts. The sequence of the first, second (and so on) letters occurred in the pile shows the secret text. The order of the pile is significant, and changing the order may yield the distinct secret text. It is easy to pile the shared natural language texts by computer operation. Human eyes can recognize the secret text from the piled texts, however, we aim to construct a natural language text secret sharing scheme employing a morphological analyzer because a meaningless phrase is a chain of morphemes consisting of one word with a high probability. We can make a shared natural text look like a natural text without any secret meaning by synthesizing using a text database.

キーワード
情報ハイディング，テキスト，ドキュメント，秘密分散，自然言語処理
Information hiding, Text, Document, Secret sharing, Natural language processing

1　まえがき

昨今、コンピュータ技術及びネットワーク技術の進歩により、デジタル形式による画像、音声、テキストなどのコンテンツの流通が爆発的に増大している。そのような状況下で、デジタルコンテンツの著作権の主張や配布先の特定、あるいは情報伝送における盗聴防止のためのカムフラージュとするために、コンテンツの中に不可視な情報を隠ぺいして埋め込む情報ハイディング技術の重要性が高まっている。

有史以来の古典的な情報ハイディングは、自然言語テキストを情報隠ぺい媒体とするものが多くを占めていた。画像における画素に対応するのは、テキストの場合は文字といえる。文字は意味の一部を成し、かつ文字と文字コードとはほぼ１対１の関係にある。したがって、テキストへ情報を隠ぺいするために文字コードへ作為を行うと、即座に文字に波及し、さらに意味にまで波及するため、テキス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、テキストを情報隠ぺい媒体とする情

報ハイディングは、テキストのレイアウト情報すなわち結局は画像の中に隠ぺいする手法が主流を占めており、プレーンテキストに適用できる手法は、一部の例外[1]-[4]を除き、あまり多く見られない。しかし、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テキス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、テキストを情報隠ぺい媒体とする情報ハイディングには、多くの応用が期待できる。

秘密分散法(secret sharing scheme)は、分散した複数の情報を合わせた場合にのみ秘匿情報を復号できる手法である[5][6]。その一つの実現形態として、Naor ら[7]によって提案された視覚復号型秘密分散法(Visual Cryptography 又は Visual Secret Sharing Scheme、以下「VSSS」という。)は、複数の半透明なスライドを重ね合わせた場合にのみ秘密画像が現れる技術であり、計算機を使わず人間の目視による復号が可能な情報隠ぺい手法として、研究や実用化が進められている[8]-[10]。

本論文では、画像コンテンツというリアルな媒体の重ね合わせによって実現される VSSS の特長に着目し、画像以外のコンテンツに同様な手法を適用する手立ての一つとして、自然言語テキストを情報隠ぺい媒体とする秘密分散法(Text Secret Sharing Scheme、以下「TSSS」という。)を提案する[11]。提案手法は、複数枚の分散テキスト(share text)を重ね合わせて、上層から下層に読んでいくと、その文字列の中に秘密テキスト(secret text)が現れるようにするものである。重ね合わせて得られる文字列から形態素解析処理によって秘密テキストを抽出する。

2 では、提案手法の原理について VSSS と対比しながら述べる。**3** では、分散テキストの生成方法と実装について説明する。**4** では、秘密テキスト抽出のための仮定の設定とその妥当性について述べる。**5** では、自然に見える分散テキストを生成するために求められる要件について議論し、提案手法の改良方策について述べる。**6** において考察を行い、**7** で今後の展望について述べる。

2 提案手法の原理

TSSS は、秘密テキストを複数の分散テキストに分散して隠ぺいし、テキストを“重ね合わせて”秘密テキストを復号するもの、と定義できる。ギリシャ時代に用いられていたとされる「スキュタレー暗号」は、ある太さのドラムに紙テープをらせん状に巻きつけ、ドラムの軸方向に文字を書き込んだ後にほどき、その紙テープを伝送し、受信者は同じ太さのドラムに巻きつけて復号するものであった。つまりドラムの太さに応じた一定間隔で文字列にスクランブルをかける暗号方式であった。TSSS において、VSSS における“重ね合わせ”に対応する処理として、このスキュタレー暗号の考え方を応用する。すなわち、複数枚の分散テキストをそれぞれ 1 行にして横書きに展開し、文字幅が均一という前提で冒頭文字の位置を合わせた際に、ある位置において縦に並んだ文字列の中に秘密テキストが現れるようにする。図 1 に例を示す。この場合、スキュタレー暗号における一巻き分の紙テープが、各分散テキスト(各行)に相当することになる。この例において、各分散テキストは、それだけで自己完結した意味を持っていることに注意する。

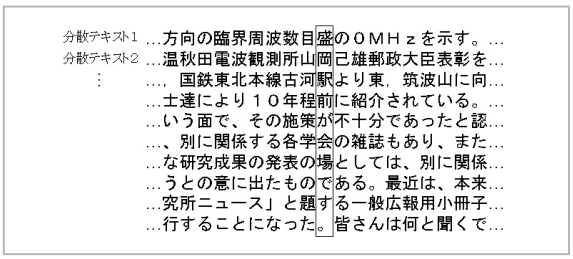


図1 TSSS の原理

重ね合わせて得られる文字列は、無意味な文字列の一部に、意味のある秘密テキストが混じっている形になる。図 1 の場合、重ね合わせて得られる文字列は、縦に読むことで、例えば、“波測線 0 のる発た」な”、“数所古年施各表もとっ”、“目山河程策学のの題た”、“盛岡駅前が会場です。”、“の己よに不のとある皆”、“0 雄り紹十雑しる一さ”などとなる。この場合の秘密テキストは「盛岡駅前が会場です。」となる。秘密テキストの探索には、自然言語の意味を利用する。つ

まり、秘密テキストとそれ以外の違いは、自然言語としての意味の有無の違いによって認識される。

以上の原理は、分散テキストの冒頭から秘密テキストを構成する各文字までの文字数を、すべての分散テキストにわたって同一にそろえておくこととし、そのことを抽出時の手がかりとするものであった。一般的には、秘密テキストを構成する各文字の、分散テキスト内での位置に関する情報を抽出時に利用できるのであれば、文字数を同一にそろえておく必然性はない。しかしそろえない場合、位置に関する情報すなわち鍵を埋め込み時と抽出時に共有しているという制約が必要になる。本論文の提案手法では、VSSS にならって、鍵を必要とする場面を少なくすることを優先するため、一つの例として、分散テキストの冒頭から秘密テキストを構成する各文字までの文字数を同一にそろえておく方法を用いている。

Shamir、Blakley らの方法において、分散データは単なるランダムなビット列であり、データはハードディスク等の記録媒体に保持することになる。分散データそのものがランダムであることから、何らかの秘密情報が含まれていることが、記録媒体にアクセス可能な第三者に明確に見破られてしまう恐れがある。特にテキストの場合、無意味な文字列が並んでいる文章はまれであり、その場合は何らかの情報が隠されているという疑念を直ちに招く懸念があり、それは秘匿に対する脅威となりえる。そのため分散テキストは、意味を持つ自然な文章になるようにし、攻撃者に疑念を抱かせない工夫を講じる必要がある。

以上の原理に基づき、**3** では、分散テキストの生成方法と実装について説明し、**4** では、秘密テキスト抽出のための仮定の設定とその妥当性について述べる。

3　分散テキストの生成

前章で述べたように、分散テキストは意味を持つ自然な文章になっていることを目指す。単語を組み合わせることによって意味を持つ自然な文章を合成することは、計算機では膨大な知識と複雑なアルゴリズムを必要とする。また、提案手法では分散テキストの文面自体は情報伝達の役目を果たす必要がないため、自然な文章に見せかけるためだけに単語からのテキスト合成を行うのは無駄な処理である。そこで、単語単位でなく文（句点）単位のテキストを大量にデータベース化しておき、そのテキストをつなぎ合わせることによって分散テキストを生成する方法を提案する。

以下に、分散テキストを生成するためのアルゴリズムを示す。

【定義】

｛ ｝は集合、＜ ＞は順序付き集合とする。英大文字は 1 テキスト（もしくはテキストの集合）、英小文字（添字を除く）は 1 文字を表す。

秘密テキスト E とは、文字列＜$e_1, e_2, \cdots, e_\varepsilon$＞から成る順序集合と定義する。ここで文字 e_i は、日本語文字であり、E は長さが ε の日本語の文とする。図 1 の例の場合、

E＝＜盛, 岡, 駅, 前, が, 会, 場, で, す, 。＞

となる。また、この例の場合、 ε は 10 となる。

テキストデータベース D は、テキストの集合 $\{T_1, T_2, \ldots\}$ とする。

D の一要素 T_x は、文字列＜$t_{x_1}, t_{x_2}, \ldots$＞から成る順序集合とする。T_x の最終要素(最後尾の文字)は、句点“。”である。T_x の長さには制限はない。

以下では、 ε が分散テキストの数と同一である場合を考える。

【処理手順】

⑴　秘密テキストの各文字を含むテキストをデータベースから抽出

$1 \leqq i \leqq \varepsilon$ であるすべての i について、文字 e_i を要素に含むテキスト T_{x_i} を D から抽出する。そして、$e_i = t_{x_i z_i}$ $(\in T_{x_i})$ とする。

その結果、

$$＜e_1, e_2, \cdots, e_\varepsilon＞ \equiv ＜t_{x_1 z_1}, t_{x_2 z_2}, \cdots, t_{x_\varepsilon z_\varepsilon}＞$$

となる。

⑵　文字数合わせ処理

次に、

$$T_{w_i} = ＜t_{w_i 1}, t_{w_i 2}, \ldots, t_{w_i y_i}＞$$

であり、

$$y_1 + z_1 = y_2 + z_2 = \ldots = y_\varepsilon + z_\varepsilon$$

（そして、この値を j とする。）

を満たす $\{T_{w_i}\}$ ($\in$D, $1 \leqq i \leqq \varepsilon$) を、D から抽出する。ただし、各 T_{w_i} は一文又はそれ以上とする。

(3) 分散テキストの合成

分散テキスト$<S_1, S_2, \ldots, S_\varepsilon>$を、

$S_1 = <T_{w_1}, T_{x_1}>$
$= <t_{w_1}1, t_{w_1}2, \ldots, t_{w_1}y_1, t_{x_1}1, t_{x_1}2, \ldots>$
$S_2 = <T_{w_2}, T_{x_2}>$
$= <t_{w_2}1, t_{w_2}2, \ldots, t_{w_2}y_2, t_{x_2}1, t_{x_2}2, \ldots>$
…
$S_\varepsilon = <T_{w_\varepsilon}, T_{x_\varepsilon}>$
$= <t_{w_\varepsilon}1, t_{w_\varepsilon}2, \ldots, t_{w_\varepsilon}y_\varepsilon, t_{x_\varepsilon}1, t_{x_\varepsilon}2, \ldots>$

とする。

(処理終)

上記の処理の原理は、$1 \leqq i \leqq \varepsilon$ のすべての i について、S_i の j 番目の文字が e_i ($= t_{x_i z_i}$) となるように、$\{S_i\}$ を合成するものである。$\{T_{w_i}\}$ は、文字数合わせのためにのみ挿入されることになる。

以上の処理手順を、perl スクリプトで実装した。テキストデータベースとしては、通信総合研究所の広報紙「CRLニュース」の 20 年分の全記事[12]を使用した。同記事をデータベースとして採用したのは、ほぼ単一分野の技術的な内容、つまり通信技術に特化された内容のテキストで統一されていて扱いやすかったためである。データベースのサイズは約 5MB である。

「盛岡駅前が会場です。」($\varepsilon = 10$) を秘密テキストとした場合に生成された分散テキスト (10枚) のうち、例として 2 枚を以下に示す。秘密テキストを構成する文字(秘密文字と呼ぶことにする) のうち、前者に「盛」、後者に「が」が含まれている。

「最近は、本来の電離層を介する伝搬よりも、むしろ宇宙通信に対して電離層が与える影響に関する研究の方が活発になっている傾向がある。もっとも内側の太線の円は衛星軌道を地球上に投影したものを表わすと同時に半径方向の臨界周波数目盛の 0MHz を示す。」

「この現象は雷放電による電波が電離層上部の多種類のイオンと作用し、特に重水素イオンと共鳴作用をすることによって生じたものと考え、これを重水素ホイッスラと呼ぶことにした。卒直に言って、当所は一般への PR という面で、その施策が不十分であったと認めざるを得ない現況である。」

アルゴリズムの動作の具体例として、秘密文字「盛」が含まれる上記の分散テキストを生成する過程を以下に示す。

(ステップ 1) テキストデータベースの中から秘密文字「盛」を含む文を探し、以下の文 T_{x_i} を抽出する (T_{x_i} を抽出文と呼ぶことにする)。

「もっとも内側の太線の円は衛星軌道を地球上に投影したものを表わすと同時に半径方向の臨界周波数目盛の 0MHz を示す。」

ただし、同じ秘密文字を含む文がデータベースに複数存在する場合は、データベース内の登録順で最初に該当する文を選択して抽出文とする。秘密文字を含む文がデータベースに存在しない場合は処理に失敗し、終了する。失敗した場合は、秘密テキストを変えるなどの対応を手動で行い、処理をやり直す。

(ステップ 2) 他の秘密文字についても同様に抽出文を探す。

(ステップ 3) すべての秘密文字について抽出文が得られたら、抽出文の中で、文頭から秘密文字までの文字数が最も多いものを探し、その文字数 (最大文字数と呼ぶことにする) をカウントする。上掲の例の場合、最大文字数は秘密文字「岡」の場合の抽出文で、110 字となる。

(ステップ 4) 残りの抽出文について、それぞれに対し「最大文字数 − 文頭から秘密文字までの文字数」を計算する。「盛」を含む抽出文の場合、文頭から秘密文字までの文字数は 47 文字となるので、「最大文字数 − 文頭から秘密文字までの文字数」は 63 文字となる。

(ステップ 5)「最大文字数 − 文頭から秘密文字までの文字数」の長さの文 T_{w_i} をテキストデータベースから抽出する。同じ長さの文がデータベースに複数存在する場合は、データベース内の登録順で最初に該当する文を選択する。「盛」を含む抽出文の場合、63 文字の T_{w_i} として、以下を抽出する。

「最近は、本来の電離層を介する伝搬よりも、むしろ宇宙通信に対して電離層が与える影響に関する研究の方が活発になっている傾向がある。」

(ステップ 6) T_{w_i} と T_{x_i} を結合したものを分散テキスト S_i とする。

以上

4　秘密テキストの抽出

分散テキストを重ね合わせて秘密テキストを抽出するためには、隠ぺいされている位置を同定する必要がある。隠ぺいされている位置に関する情報を抽出時に利用できない場合には、秘密テキストとそうでない文字列との切り分けを自然言語処理的に行う必要がある。そのためには、秘密テキストが意味を持つフレーズであるとする制約が必要となる。この制約は、VSSS において、無意味な画像を秘密原画像とした場合には、輪郭が判別できないため、背景から秘密原画像を切り出すことが目視では困難であることと同等であるので、秘密分散法の応用としては妥当な制約である。この制約では、分散テキストを重ね合わせた際に、秘密テキスト以外の箇所が偶然に意味を持つフレーズになる可能性は小さいと考えられるため、自然言語処理により、意味を持つフレーズを抽出することが、秘密テキストを抽出することと等価であるとみなすことができる。したがって、秘密テキストは目視によっても抽出できることになるが、本論文では自然言語処理を援用することによる抽出方法を考える。

上記の前提に基づき、自然言語処理による秘密テキスト抽出のために、以下の仮定を置く。

[仮定]

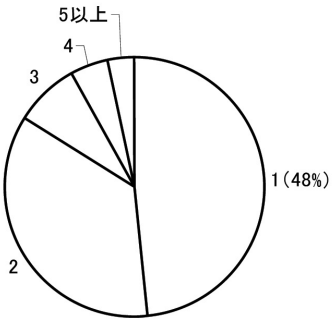
自然言語処理における基本的な処理の一つとして、テキストを形態素(語を構成する最小単位)に分解する形態素解析がある。形態素解析をすると、意味を持たないフレーズは、1 文字の形態素の連鎖になる割合が多い。

上記の仮定の妥当性を検証するために、意味のある文と、ランダムな文字列とをそれぞれ形態素解析し、1 文字形態素の連鎖の出現頻度を比較した。

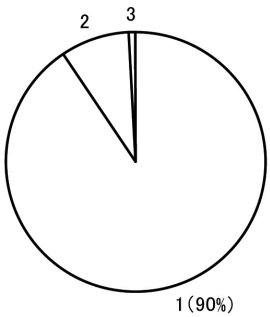
学術解説文[13]の本文のみを取り出した、意味を持つ 8098 個の日本語文字列と、同じ文章を基に生成したランダム文字列*とをそれぞれ形態素解析した。ここで形態素解析器として「茶筌」[15]を用い、形態素辞書として同解析器の標準添付辞書を用いた。前者の形態素数は 4444 個、後者は 7062 個となった。それぞれについて、形態素の文字数と出現比率との関係を図 2 に示す。学術解説文の場合、1 文字形態素が全体の半分弱(48%)であったのに対し、ランダム文字列の場合は全体の 90 %を占め、最長の形態素でも 3 文字までであった。次に、1 文字形態素の連鎖の長さと出現比率との関係を図 3 に示す。学術解説文では、図 2 に示したとおり、1 文字形態素は全体の半分弱を占めていたにもかかわらず、3 個以上の連鎖になっていたのは全体の 11 %に過ぎなかったのに対し、ランダム文字列の場合は全体の 86 %を占めていた。以上の実験結果より、1 文字形態素の連鎖の短い個所を手掛かりとして、意味を持つフレーズを高い精度で抽出できると結論でき、仮定の妥当性が示されたといえる。そこで、実装においては、茶筌の出力に対して、1 文字形態素の連鎖が 3 個未満であるフレーズを、秘密テキストの候補とする。この閾値の場合、ランダム文字列を秘密テキストとして抽出する誤り率、すなわち「1－適合率」は 14 %となり、意味を持つフレーズを見落とす誤り率、すなわち「1－再現率」は 11 %と見積もることができる。閾値として、1 文字形態素の連鎖を長くすると、再現率は高くなるが適合率は低くなり、連鎖を短くすると、その逆となる。

図 1 に示した分散テキストについて、正しい順序に重ね合わせて形態素解析した結果の一部を図 4 に示す。1 文字形態素の連鎖が 3 個未満のフレーズである「盛岡駅前が会場です。」(括弧で示した部分)が、秘密テキストの候補として抽出されることになる。

＊　文字列を暗号化/復号するフリーソフト BookNoise ver1.01[14]を用い、文献[13]をいったん暗号化(ASCII 文字列化)し、別の鍵を用いて全く別の日本語文字列として復号することで、ランダム化したとみなした。したがって数学的に厳密なランダム文字列ではない。



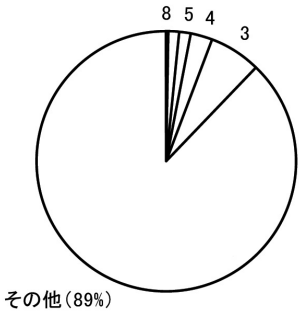
学術解説文
meaningful text



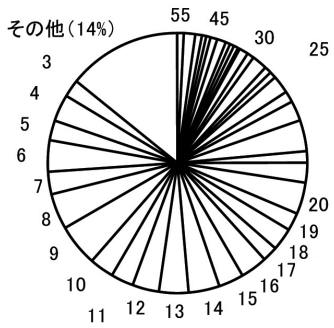
ランダム文字列
random character sequence

図2 形態素の文字数と出現比率の関係

（数字は形態素の文字数）



学術解説文
meaningful text



ランダム文字列
random character sequence

図3 1文字形態素の連鎖の長さと出現比率の関係

（数字は1文字形態素の連鎖の長さ）

5 分散テキストの自然性についての議論

秘密分散法では、分散データが欠落すると秘密情報を完全には復号できなくなるため、分散データに情報が隠ぺいされていることを第三者に見破られて破棄あるいは改ざんされることが脅威となる。見破られないようにするためには、生成された分散テキストの自然性が保たれていることが重要である。

提案手法には、文自体を合成するプロセスはなく、分散テキストは、人間が作成した既存の文をテキストデータベースから複数個取り出してそのまま並べる処理だけで作成される。そのため提案手法においては、各文の単体での自然性については問題はなく、複数並べられた各文間のつながりの自然性をどのように保つかが課題となる。

各文間のつながりの自然性は、各文がある話題について一貫しているかどうか、すなわち意味的にまとまっているかどうかによって決まる。山本らは、文章の意味的なまとまりが、文同士の「結束性」(cohesion)によって定量化できることを示している[16]。山本らによると、結束性は、

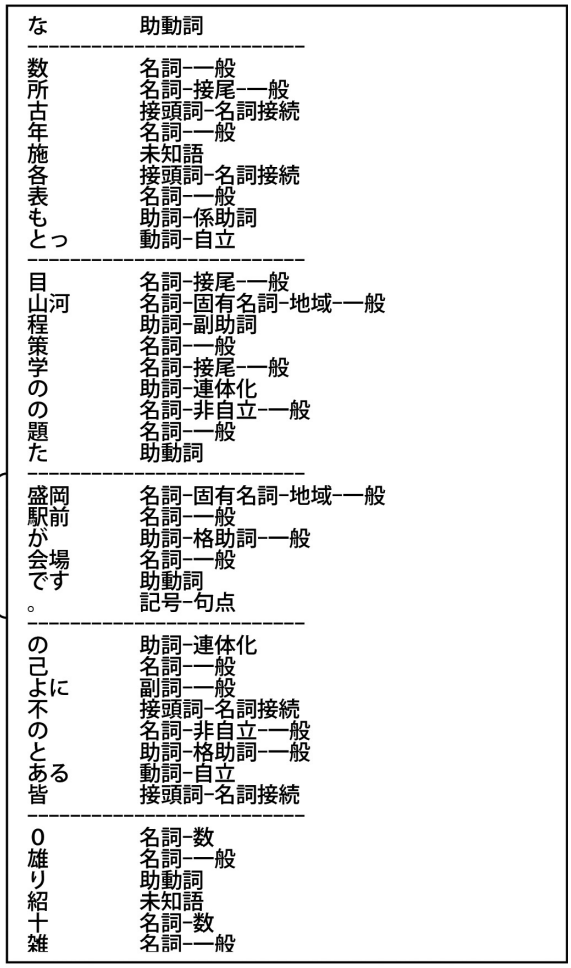


図4　図１の分散テキストを形態素解析した結果

接続詞や副詞などの接続的語句と、類似した意味を持つ語句（同一語句、上位・下位語、類義語、対義語）とで表現される。すなわち、2 文が接続的語句で接続されていたり、類似した意味を持つ語句の出現頻度が高かったりする場合に、結束性が強いことになる。そこで、提案手法において、データベースから抽出する文の候補が複数存在する場合に、データベース内の登録順で最初に該当する文を選択していた処理について、以下の処理で置き換えることにより、分散テキストの自然性を向上させることができる。

秘密テキストを構成する文字（秘密文字）を含む文 T_{x_i} 及び $1 \leqq i \leqq \varepsilon$ にわたって冒頭からの文字数をそろえるための文 T_{w_i} について、山本らの手法に基づいて T_{x_i} と T_{w_i} の結束性を計算し、$1 \leqq i \leqq \varepsilon$ にわたっての結束性の総和が最大となる T_{x_i} と T_{w_i} の組合せを、分散テキストとして選ぶ。

提案手法では、各分散テキストの内容は任意であり、文章としての自然性さえ保たれていればよいので、テキストデータベースについては内容の制約なく任意に採用することができる。したがって、上記の処理による結束性の向上によって自然性を向上できるほか、手法はそのままでテキストデータベースのみを置き換えることによって、更に分散テキストの自然性を向上させることができる。例えばジャンルやカテゴリが細分化された文データから成るテキストデータベースを使い、T_{x_i} と T_{w_i} は同じ細分類に属する文の組合せに限定する、といった制約を設けることにより、生成される分散テキストの自然性をより向上させることができる。

このように提案手法は、分散テキストの自然性を向上させるための拡張が可能な手法であるといえる。

6 考察

重ね合わせる順序を鍵とする一部の方式[9]を除き、VSSSの多くは、秘密データを抽出する際に分散データを重ね合わせるだけで鍵を使用しないシンプルな方法である。それに対し提案手法は、重ね合わせる順序を鍵として、埋め込み時と抽出時にこの鍵を共有していることを前提としている。しかしながら、重ね合わせる順序が異なれば抽出処理によって意味を持つフレーズを全く見つけられないことになるため、原理的には、重ね合わせ順序についての情報を用いず、順序を総当たりで入れ替えて解析したとしても、秘密テキストを正しく抽出できる原理になっている。重ね合わせる順序はε！通りあるため、εが大きいと総当たりにかかる計算量の増大が問題となるが、形態素解析の高速化や並列処理などの手立てを講じることで、VSSSと同様に、鍵を用いない抽出法の実現も可能な手法といえる。

提案手法は、(秘密テキストの文字数)≦(分散テキストの枚数)である必要があり、長い秘密テキストを埋め込むためには多数の分散テキストが必要になる。したがって提案手法は、あまり長い秘密テキストを用いる用途には適していない。秘密情報に関するこの制約は、利用法に制約を課す性質であるが、VSSSにおいて、細かい画像やコントラストの低い画像を秘密画像として扱うことが難しい制約と同等であり、秘密分散法をコンテンツに適用するためにはやむを得ない制約といえる。秘密分散法は、秘密情報を分散することにより、鍵回復や鍵供託への応用があり、マルチパーティープロトコルを実現する際の基本的な構成要素でもある。さらに暗号通信を行う際に、複数の暗号技術を利用して安全性を高める暗号の多重化にも利用することができる。それに対して提案手法は、前述のとおり長い秘密テキストの分散には不向きなため、鍵の分散共有法としての応用が適しているといえる。

一般の秘密分散法は、分散データが一つでも欠落していれば、秘密データを全く復号できない特徴を持っている。それに対し、提案したテキスト秘密分散法の場合、分散テキストの欠落が減っていくにつれて秘密テキストが徐々に完成していくという性質がある。そのため、秘密テキストの欠落文字を文脈に基づいて補間できる程度まで分散テキストがそろっていれば、完全にそろっていなくても事実上の復号ができてしまう性質がある。これは分散データと秘密データを共に意味を持つ自然言語テキストとしている限りにおいて不可避な性質である。ただし、秘密テキストを形態素解析によって抽出する提案手法では、一部の秘密文字が欠落している場合には形態素が分断される可能性が高いため、一部欠落した秘密テキストは意味を持つフレーズとしての抽出に失敗し、結果的に復号を回避できる率が高いことが期待される。分散テキストの欠落の程度と、秘密テキストの復号に関する安全性との関係については、今後定量的に検証する必要がある。

7 むすび

本論文では、自然言語テキストを情報隠ぺい媒体とする秘密分散法を提案し、分散テキストの生成機能と、秘密テキストの抽出機能を実装した結果について述べた。秘密テキストの抽出機能については、意味のある文字列とランダム文字列について形態素解析を行った結果を比較し、1文字形態素の連鎖が3個未満となるフレーズを閾値として設定することが妥当であることを示した。また、生成した分散テキストの自然性について議論し、分散テキストを構成する文間の結束性の概念を用いることによって、提案手法を改良できることを示唆した。

今後は、考察において指摘した課題の解決を中心に進め、主観的評価実験を含めて、生成した分散テキストの自然性についての評価を行う。また、提案手法に適したテキストデータベースのあり方について検討し、多くの秘密テキストに適用して評価を行う。さらに、提案手法の利用法についての検討も進める。

謝辞

本研究のきっかけを与えてくださった、PuKyong National University の故 Ji-Hwan Park 教授に感謝する。また、自然言語テキストを情報隠ぺい媒体として適用する方法に関して、横浜国立大学の松本勉教授及び松本研究室の諸氏、東京大学の中川裕志教授、株式会社三菱総合研究所の諸氏から有益な助言を賜っていることに感謝する。

参考文献

1 Mikhail Atallah, et al., "Natural Language Watermarking and Tamperproofing", 5th International workshop on information hiding, Oct. 2002.
2 中川裕志，木村浩康，三瓶光司，松本勉，“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.8，2272- 2279，2000.
3 松本勉，中川裕志，村瀬一郎，“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.
4 滝澤修，“情報埋込・抽出方法及びその装置並びに記録媒体”，特開 2002-269074.
5 A.Shamir, "How to share a secret", Communications of the ACM, pp.612-613, 1979.
6 G.Blakley, "Safeguading cryptographic keys", Proceedings of AFIPS National Computer Conference, pp.313-317, 1979.
7 M.Naor and A.Shamir, "Visual Cryptography", Advances in Cryptology-Eurocrypt'94, pp.1-12, 1994.
8 加藤拓，今井秀樹，“視覚復号型秘密分散法の拡張構成方式”，電子情報通信学会論文誌，Vol.J79-A，No.8，pp.1344-1351，1996 年 8 月.
9 有井幸太，盛拓生，坂井一雄，今井秀樹：“積み重ね順序を鍵とする視覚暗号方式”，SCIS2000，2000 年 1 月.
10 視覚復号型暗号製品「あわすとで～る」，凸版印刷株式会社，http://www.toppan.co.jp/aboutus/release/article463.html，2001 年 4 月.
11 滝澤修，山村明弘，“自然言語テキストを用いた秘密分散法”，情報処理学会論文誌，Vol.45，No.1，pp.320-323，2004 年 1 月.
12 通信総合研究所，“CRLニュース”，創刊号～第 238 号，1976 年～1995 年.
13 中川裕志，滝澤修，井上信吾，“ドキュメントへのインフォメーションハイディング”，情報処理，Vol.44，No.3，pp.248-253，2003年3月.
14 "BookNoise，ver.1.01", http://www.vector.co.jp/soft/win95/util/se267011.html
15 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），“日本語形態素解析システム茶筌 version 2.0 for Windows”，1999.
16 山本和英，増山繁，内藤昭三，“段落分けを用いた日本語文章における結束構造の検討”，情報処理学会論文誌，Vol.35，No.10，pp.2029-2037，1994年10月.

滝澤　修（たきざわ　おさむ）
情報通信部門セキュリティ高度化グループ主任研究員　博士（工学）
コンテンツセキュリティ、非常時防災通信

山村明弘（やまむらあきひろ）
情報通信部門セキュリティ基盤グループリーダー　Ph.D.
暗号理論、情報セキュリティ

牧野京子（まき のきょうこ）
株式会社三菱総合研究所
ソフトウェア工学