テクニカルノート

自然言語テキストを用いた秘密分散法

滝澤 修[†] 山村明弘[†]

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新しい秘密分散法を提案する．提案手法は，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．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 1 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．

Secret Sharing Scheme Using Natural Language Text

Osamu Takizawa[†] and Akihiro Yamamura[†]

Modifying the idea of the visual cryptography, we propose a method of sharing a secret key using natural language texts. 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. 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.

1. まえがき

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抽出できる手法である[1),2)]．その 1 つの実現形態として，Naor ら[3)]によって提案された視覚復号型秘密分散法（Visual Cryptography または Visual Secret Sharing Scheme，以下 VSSS）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抽出が可能な情報隠蔽手法として，研究や実用化が進められている[4)~6)]．

本論文では，画像というリアルな分散データ（share）の重ね合わせによって秘密データ（secret data）を抽出できる VSSS の特長を，画像以外のコンテンツでも実現することを目指し，自然言語テキストを分散データとする新しい秘密分散法（Text Secret Sharing Scheme，以下 TSSS）を提案する．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. 提案手法の原理

TSSS では，VSSS における “重ね合わせ” に対応する処理として，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．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（秘密文字と呼ぶことにする）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揃えておくことによって，秘密文字の，分散テキスト内における位置に関する情報（これを鍵と定義する）を，埋め込み時と抽出時に共有する必要がない．したがって，鍵を使わずに重ね合わせだけで抽出できる VSSS と同様なシンプルさを実現できることになる．

複数個の分散テキストをそれぞれ 1 行ずつにして横書きに展開した例を**図 1** に示す．重ね合わせて得られる文字列を縦に読み，秘密テキストは「盛岡駅前が会場です。」となる．

† 独立行政法人通信総合研究所
Communications Research Laboratory

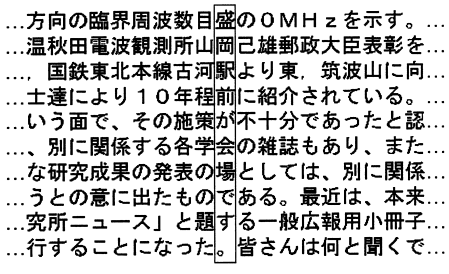

図 1 TSSS の原理

Fig. 1 A concept image of the proposed text secret sharing scheme.

3. 分散テキストの生成

無意味な文字列が並んでいるテキストは，何らかの秘匿情報が含まれているという疑念をただちに招く懸念がある．そのため分散テキストは，意味を持つ自然な文章になる工夫を講じる必要がある．そこで，文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採用する．

【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 $\langle e_1, e_2, \cdots, e_\varepsilon \rangle$ 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，

$E = \langle$ 盛，岡，駅，前，が，会，場，で，す，。$\rangle$

となる．また，この例の場合，ε は 10 となる．本論文では，ε が分散テキストの数と同一である場合を考える．

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 $\langle t_{x1}, t_{x2}, \ldots \rangle$ から成る順序集合とする．

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．また具体例として，「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，秘密文字である「盛」が含まれている分散テキストを生成する過程を並行して示す．以下に示すアルゴリズムは perl スクリプトで実装した．またテキストデータベースとして，通信総合研究所の広報紙「CRL ニュース」の 20 年分の全記事[7]を使用した．データベースのサイズは約 5 MB であった．

【処理手順】

1．秘密文字を含むテキストをデータベースから抽出

$1 \leq i \leq \varepsilon$ 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 から選ぶ（T_{x_i} を抽出文と呼ぶことにする）．

そして，$e_i = t_{x_i z_i} (\in T_{x_i})$ とする．

その結果，

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

となる．

例の場合，テキストデータベースで秘密文字 $e_i =$「盛」を含む文を検索した結果，以下の文 T_{x_i} が選ばれる．

「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 0 MHz 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．

2．文字数合わせ処理

次に，

$$T_{w_i} = \langle t_{w_i 1}, t_{w_i 2}, \ldots, t_{w_i y_i} \rangle$$

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 $\{T_{w_i}\} (\in D, 1 \leq i \leq \varepsilon)$ を，D から選ぶ．ただし各 T_{w_i} は 1 文またはそれ以上とする．

具体的にはまず，すべての秘密文字についての抽出文の中で，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．例の場合，最大文字数は秘密文字「岡」の場合で，110 字となる．その他の抽出文について，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」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，T_{w_i} として，63 文字である以下の文が D から選ばれる．

「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。」

3．分散テキストの合成

分散テキスト $\langle S_1, S_2, \ldots, S_\varepsilon \rangle$ を，

$$
\begin{aligned}
S_1 &= \langle T_{w_1}, T_{x_1} \rangle \\
&= \langle t_{w_1 2}, t_{w_1 2}, \ldots, t_{w_1 y_1}, t_{x_1 1}, t_{x_1 2}, \ldots \rangle \\
S_2 &= \langle T_{w_2}, T_{x_2} \rangle \\
&= \langle t_{w_2 1}, t_{w_2 2}, \ldots, t_{w_2 y_2}, t_{x_2 1}, t_{x_2 2}, \ldots \rangle \\
&\ldots \\
S_\varepsilon &= \langle T_{w_\varepsilon}, T_{x_\varepsilon} \rangle \\
&= \langle t_{w_\varepsilon 1}, t_{w_\varepsilon 2}, \ldots, t_{w_\varepsilon y_\varepsilon}, t_{x_\varepsilon 1}, t_{x_\varepsilon 2}, \ldots \rangle
\end{aligned}
$$

とする．そして，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．

「盛」の S_i は以下のとおりとなる．

「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すと同時に半径方向の臨界周波数目盛の 0 MHz を示す。」

（処理終）

以上の処理の原理は，$1 \leq i \leq \varepsilon$ 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，$\{Si\}$ 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．

4. 秘密テキストの抽出

提案手法は，秘密文字の位置に関する情報を秘密テキストの抽出に利用せず，自然言語処理によって抽出する．そのため，秘密テキストが意味を持つ文字列であるとする制約を設ける．この制約は，VSSS において，目視による抽出が困難である無意味な画像を秘密原画像とはできないことと同等であるので，妥当な制約と考えられる．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文字列になる可能性は小さいと見なせるため，自然言語処理により，意味を持つ文字列を抽出することが，秘密テキストを抽出することと等価であると見なせる．ただし本論文では自然言語処理を援用することによる抽出を考えるが，原理的には VSSS と同じく目視によっても抽出できる．

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 1 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，意味を持たない文字列は，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[8]の本文のみを取り出した，意味を持つ 8,098 個の日本語文字列と，同じ文章を元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」[10]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 4,444 個，後者は 7,062 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．**図 3** に，1 文字形態素の連鎖の長さと出現比率との関係を示す．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%にす

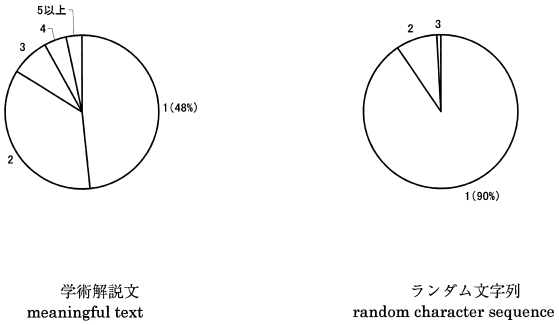

図 2 形態素の文字数と出現比率の関係（数字は形態素の文字数）
Fig. 2 Relation between number of characters of morpheme and appearance ratio.

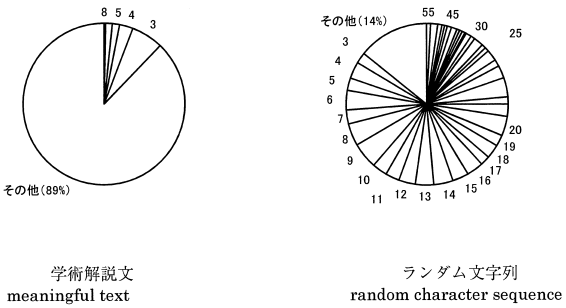

図 3 1 文字形態素の連鎖の長さと出現比率の関係（数字は 1 文字形態素の連鎖の長さ）
Fig. 3 Relation between chain length of 1 character morpheme and appearance ratio.

☆ 文字列を暗号化/復号するフリーソフト BookNoise ver1.01[9] を用い，学術解説文[8]をいったん暗号化（ASCII 文字列化）し，別の鍵を用いてまったく別の日本語文字列として復号することで，ランダム化したと見なした．したがって，数学的に厳密なランダム文字列ではない．

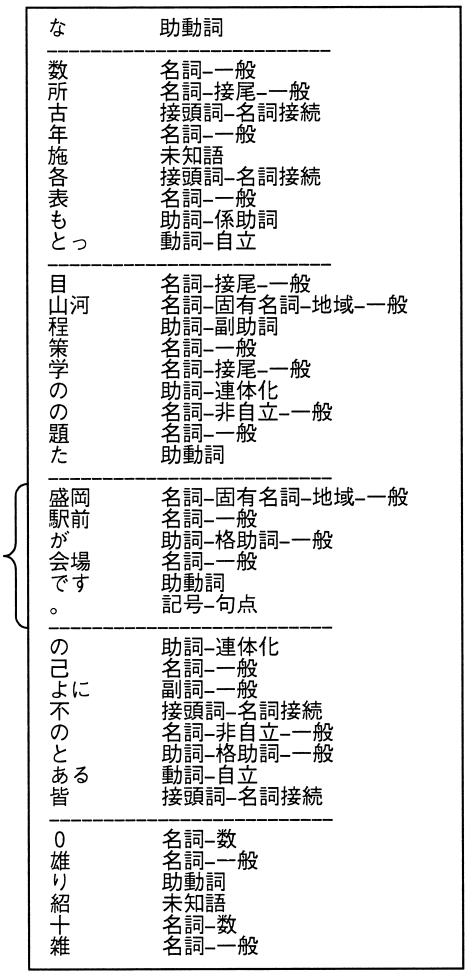

図 4 図 1 の分散テキストを形態素解析した結果（一部）
Fig. 4 A part of the result of morphological analysis of the example text shown in Fig. 1.

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箇所を手がかりとして，意味を持つ文字列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 個未満である文字列を，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，意味を持つ文字列を見落とす誤り率，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 個未満の文字列である「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．

5. むすび

本論文では，自然言語テキストを分散データとする新しい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．

今後は，生成した分散テキストの自然性についての評価，提案手法に適したテキストデータベースの構築方法，提案手法の利用法についての検討などが必要である．

謝辞 本研究のきっかけを与えてくださった，Pu-Kyong National University の Prof. Ji-Hwan Park に感謝する．また，横浜国立大学の松本勉教授および松本研究室の諸氏，東京大学の中川裕志教授，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) Shamir, A.: How to share a secret, *Comm. ACM*, Vol.22, No.11, pp.612–613 (1979).
2) Blakley, G.: Safeguarding cryptographic keys, *Proc. AFIPS National Computer Conference*, Vol.48, pp.313–317 (1979).
3) Naor, M. and Shamir, A.: Visual Cryptography, *Advances in Cryptology-Eurocrypt'94*, pp.1–12 (1994).
4) 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).
5) 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，暗号と情報セキュリティシンポジウム，SCIS2000-B46 (2000).
6) 視覚復号型暗号製品「あわすとで〜る」，凸版印刷株式会社 (2001). http://www.toppan.co.jp/aboutus/release/article463.html
7) 通信総合研究所：CRL ニュース，創刊号〜第 238 号 (1976–1995).
8) 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).
9) BookNoise ver.1.01. http://www.vector.co.jp/soft/win95/util/se267011.html
10) 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).

(平成 15 年 7 月 8 日受付)
(平成 15 年 10 月 16 日採録)