Method of Hiding Information in Agglutinative Language Documents Using Adjustment to New Line Positions

Osamu Takizawa[1], Kyoko Makino[2], Tsutomu Matsumoto[3],
Hiroshi Nakagawa[4], and Ichiro Murase[2]

[1] National Institute of Information and Communications Technology
4-2-1, Nukuikita-machi, Koganei, Tokyo 184-8795, Japan
taki@nict.go.jp
[2] Mitsubishi Research Institute Inc.
3-6, Otemachi 2-chome, Chiyoda, Tokyo 100- 8141, Japan
[3] Graduate School of Environment and Information Sciences,
Yokohama National University
79-7, Tokiwadai, Hodogaya, Yokohama 240- 8501, Japan
[4] Information Technology Center,
University of Tokyo
7-3-1, Hongo, Bunkyo, Tokyo 113-0033, Japan

Abstract. Information hiding technology embeds information using the redundancy of information contained in cover data. Therefore, many information-hiding techniques for cover data with a lot of redundancy, such as images or sound signals, have been proposed. Most proposed information hiding techniques that set document to cover data tampered with layouts between spacing and words. In this paper, a new information hiding technique for agglutinative languages like Japanese or Korean that have no spaces between morphemes is proposed. By the proposed technique, digital documents are set to cover data and secret data is embedded by making the position of the new-line code inserted into document correspond to secret data. The technique can also be applied to plain text like an e-mail, which does not have layout information. Because the technique does not change the content of the cover data at all, the technique can be used not only as steganography aiming at performing secret communication but also as digital watermarking. Moreover, the technique has the feature whereby embedded data remains also in the printing output.

1 Introduction

Information hiding technology embeds information using the redundancy of information contained in cover data. Therefore, many information hiding techniques for cover data with a lot of redundancy, such as images or sound signals, have been proposed. On the other hand, with information hiding that sets document to cover data, secret data is embedded into cover data, i.e., cover text, and it is set to stego text. There is no redundancy in the character code of cover text, so it is difficult to embed secret data using character code. Therefore, with information hiding that sets document to cover data, there are many techniques of considering image documents, tampering with a layout, and embedding secret data [1]. In tampering with the layout, a

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1039–1048, 2005.

slight expansion or reduction of a line interval, a word interval, or character width, and slight rotation of a character are proposed. For example, if the number of standard pixels of a line interval is defined, the interval will be expanded in order to embed a bit “1”, and the interval will be narrowed in order to embed a bit “0”. Extraction of the embedded data will be performed by detecting expansion or reduction of the line interval by using a scanner. Therefore, the extraction success rate of embedded data will depend on the scanner resolution. If the grade of expansion or reduction is made small, an extraction error will increase instead of mental-fatigue-coming to be hard to tampering more.

On the other hand, the following researches exist for the technique of embedding secret data by character code into document instead of layout.

(1) SNOW
A technique of embedding secret data by setting English as the cover text and inserting space characters that are not visible on printed matter or a screen at the end of a line has been proposed. The name of the technique is SNOW [2]. SNOW is the technique of embedding 3 bits of secret data per line by inserting zero to seven space characters at the end of each line corresponding to secret data. By this technique, when the output system in which space characters are disregard or displayed as space is used, deterioration of the document does not occur. However, by using a certain kind of text editor, it can be seen that many unnatural space characters exist at the end of the lines. Moreover, a machine can discover the existence of unnatural space characters easily. When stego text is edited using an editor that erases excessive space characters, secret data is lost. Moreover, since the embedded information disappears, space characters with disregard or the output by output system displayed as space, SNOW is the limited technique which can be used only in the circulation as electronic data.

(2) Using the number of words of each line in a LaTeX source file
A technique that embeds secret data by adjusting the number of words of each line of English LaTeX source file by setting the source file to cover text has been proposed [3]. This technique uses the redundancy that the display document after compiling does not have information on the number of words of each line in the source file in general LaTeX. This technique is a kind of information hiding in computer program codes rather than information hiding in a document.

(3) FinPri.txt
A technique of embedding secret data, without changing the meaning of a text a lot is proposed by replacing words in the cover text with synonyms. The name of the technique is FinPri.txt [4]. This is the technique of using synonyms as redundancy of vocabularies. This technique, whose chief aim is saving rough meaning of the document, can be used for technical writing documents, such as a manual. Moreover, since an aggressor cannot discover the method of embedding secret data easily, there is the feature strong against the attack that removes secret data. However, there is a deterioration in documents in which importance is attached to delicate nuance when synonyms have been substituted, such as literary works or contracts. Moreover, since the alteration of vocabulary is an act that infringes on copyright, except when processed

by the author him/herself, there is a possibility of being restricted by law. Therefore, documents to which this technique can be applied are limited.

In addition, NICETEXT [5] and Texto [6] have been proposed. These methods are text generators that convert secret data into trivial or nonsense text. Therefore, they can't be applied to watermarking or fingerprinting for literary works.

Syntactically based and semantically based watermarking methods for natural language text have been proposed [7][8]. These methods require sophisticated natural language processing.

Linguistic steganography for Russian has been proposed [9]. The method is similar to the FinPri.txt mentioned above. In this method, words of the source text, i.e., cover text, are replaced with their synonyms. Therefore, as the text becomes deteriorated in comparison with the source text, the method is inapplicable to literary works. Moreover, the method needs a large synonymy dictionary and a huge collocation database.

The proposed technique in this paper does not insert invisible character code, and does not tamper with layout, but inserts new-line code that does not affect a document. In an agglutinative language like Japanese or Korean, it is comparatively free to start a new line in the middle of morphemes. That is, even if it starts a new line in the middle of morphemes, a hyphenation is not required. Then, secret data is embedded by making the position at which a new-line code is inserted correspond to secret data.

2 Outline of the Proposed Technique

The concept of procedure by the proposed technique is shown in Fig. 1. In the cover text the new-line code is contained only at the end of a paragraph, like that in a word processor document, and it is referred to as digital document by which the new-line code is not contained in the right end of every line on screen or paper. And the document which is inserted the new-line code for every line is set to stego text, making it correspond with the information to be embedded according to rules defined before-

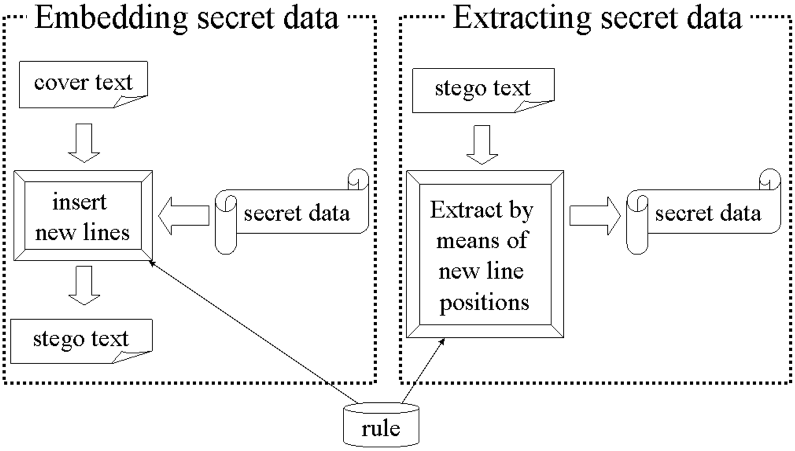
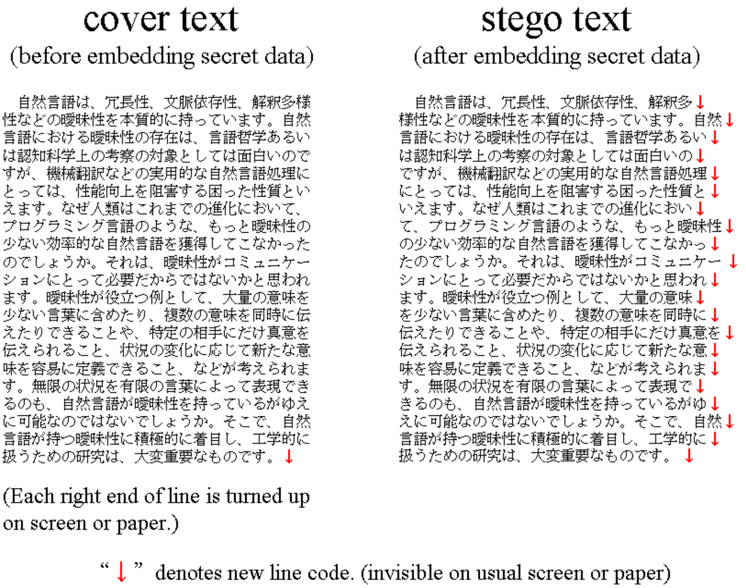


Fig. 1. Processing concept in which secret data is made to correspond to the position of a new-line code, embeds secret data into cover text, and is extracted from stego text

hand. In this technique the correspondence rule is a key for the encryption and decryption, and the same key is used for procedure both embedding and extracting secret data. This is a technique using being hard to distinguish the true new line with a new-line code from the turned up portion at right end of the line on general display system. An example of cover text and stego text is shown in Fig. 2.



cover text

(before embedding secret data)

　自然言語は、冗長性、文脈依存性、解釈多様
性などの曖昧性を本質的に持っています。自然
言語における曖昧性の存在は、言語哲学あるい
は認知科学上の考察の対象としては面白いので
すが、機械翻訳などの実用的な自然言語処理に
とっては、性能向上を阻害する困った性質とい
えます。なぜ人類はこれまでの進化において、
プログラミング言語のような、もっと曖昧性の
少ない効率的な自然言語を獲得してこなかった
のでしょうか。それは、曖昧性がコミュニケー
ションにとって必要だからではないかと思われ
ます。曖昧性が役立つ例として、大量の意味を
少ない言葉に含めたり、複数の意味を同時に伝
えたりできることや、特定の相手にだけ真意を
伝えられること、状況の変化に応じて新たな意
味を容易に定義できること、などが考えられま
す。無限の状況を有限の言葉によって表現でき
るのも、自然言語が曖昧性を持っているがゆえ
に可能なのではないでしょうか。そこで、自然
言語が持つ曖昧性に積極的に着目し、工学的に
扱うための研究は、大変重要なものです。↓

(Each right end of line is turned up on screen or paper.)

stego text

(after embedding secret data)

　自然言語は、冗長性、文脈依存性、解釈多↓
様性などの曖昧性を本質的に持っています。自然↓
言語における曖昧性の存在は、言語哲学あるい↓
は認知科学上の考察の対象としては面白いの↓
ですが、機械翻訳などの実用的な自然言語処理↓
にとっては、性能向上を阻害する困った性質と↓
いえます。なぜ人類はこれまでの進化におい↓
て、プログラミング言語のような、もっと曖昧性↓
の少ない効率的な自然言語を獲得してこなかっ↓
たのでしょうか。それは、曖昧性がコミュニケー ↓
ションにとって必要だからではないかと思われ↓
ます。曖昧性が役立つ例として、大量の意味↓
を少ない言葉に含めたり、複数の意味を同時に↓
伝えたりできることや、特定の相手にだけ真意を↓
伝えられること、状況の変化に応じて新たな意↓
味を容易に定義できること、などが考えられま↓
す。無限の状況を有限の言葉によって表現で↓
きるのも、自然言語が曖昧性を持っているがゆ↓
えに可能なのではないでしょうか。そこで、自然↓
言語が持つ曖昧性に積極的に着目し、工学的に↓
扱うための研究は、大変重要なものです。 ↓

" ↓ " denotes new line code. (invisible on usual screen or paper)

Fig. 2. Hiding information by inserting new-line codes in Japanese document. (Monospace fonts are used and each line is not justified in the stego text)

This technique has the following features.

(1) In this technique excessive codes like space characters are not inserted. The new-line code that is indispensable for documents is only used.

(2) If the output system does not disregard new-line code, the secret data remains also in printing output. This feature suggests this technique is applicable not only for electronic circulation but also for hard copy circulation.

(3) Since words are not replaced at all, the document does not deteriorate. Therefore, it can be applied to a literary work, or a contract, and it can be used as digital watermarking for it not only can be used as steganography, but also asserting the right of the work.

3 Detail of the Proposed Technique

3.1 Secret Data

Secret data is taken as a bit sequence of 0 or 1. To distinguish the line where secret data is embedded and not embedded at the stage of extraction, and to show the range of secret data, the flag sequences of the start and the end are used.

3.2 Proposal of Two Methods

The following two methods are proposed as correspondence rules.

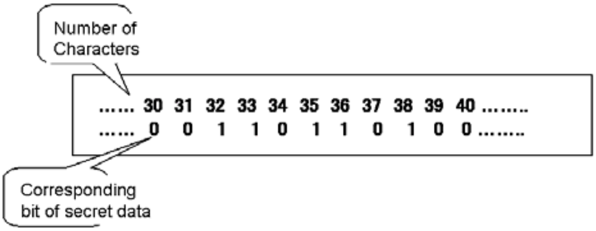
3.2.1 The Method to Make the Number of Characters per Line Correspond to Secret Data

The correspondence rule of this method defines the corresponding table of the number of characters per line, and the bit of secret data corresponding to it.

The embedding procedure is as follows. The new-line code is inserted in the position which becomes the number of characters per line corresponding to the bit of the secret data which is going to be embedded. However, priority is given to maintaining uniform line width in a general display system in case the number of characters per line is chosen from a corresponding table. Here, line width is defined as the total of the character width of all the characters in the line concerned. The Japanese character width of so-called "single-byte character" is defined as 1 per 1 character, and the character width of so-called "double-byte character" is defined as 2 per 1 character.

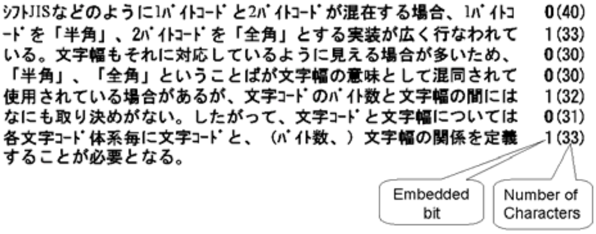
The extraction procedure is as follows. The number of characters of each line is counted and secret data is extracted using the same corresponding table.

An example of the corresponding table and an example of stego text generated by the table are shown in Figs. 3 and 4. In this system, single bit of secret data is embedded per line.



……	30	31	32	33	34	35	36	37	38	39	40	…….
……	0	0	1	1	0	1	1	0	1	0	0	…….

Fig. 3. Example of the rule table corresponding to number of characters of each line and single bit of secret data



ｼﾌﾄJISなどのように1ﾊﾞｲﾄｺｰﾄﾞと2ﾊﾞｲﾄｺｰﾄﾞが混在する場合、1ﾊﾞｲﾄｺ 0(40)
ｰﾄﾞを「半角」、2ﾊﾞｲﾄｺｰﾄﾞを「全角」とする実装が広く行なわれて 1(33)
いる。文字幅もそれに対応しているように見える場合が多いため、 0(30)
「半角」、「全角」ということばが文字幅の意味として混同されて 0(30)
使用されている場合があるが、文字ｺｰﾄﾞのﾊﾞｲﾄ数と文字幅の間には 1(32)
なにも取り決めがない。したがって、文字ｺｰﾄﾞと文字幅については 0(31)
各文字ｺｰﾄﾞ体系毎に文字ｺｰﾄﾞと、（ﾊﾞｲﾄ数、）文字幅の関係を定義 1(33)
することが必要となる。

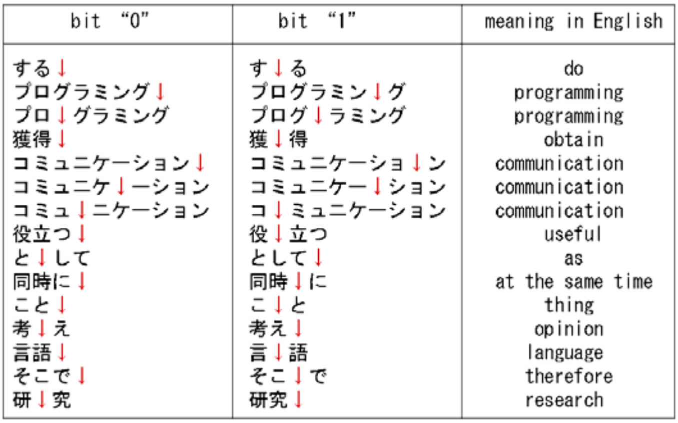
Embedded bit
Number of Characters

Fig. 4. Example of generated and justified stego text and embedded bits using the corresponding table shown in Fig. 3

The same corresponding table is used in procedure both embedding and extraction of secret data. Moreover, standard line width and minimum line width are defined, and they are used in procedure both embedding and extraction of secret data. Standard line width is a parameter for specifying standard line width to make the document of a favorite layout. Minimum line width is a parameter for not embedding bits of secret data in lines under the specified line width. It is made for line width not to embed secret data in a remarkable short line like the last line of a paragraph, or a caption by specifying minimum line width. Minimum line width serves as a required parameter in procedure of embedding and extraction.

3.2.2 The Method to Make New-Line Position in Morpheme Correspond to Secret Data

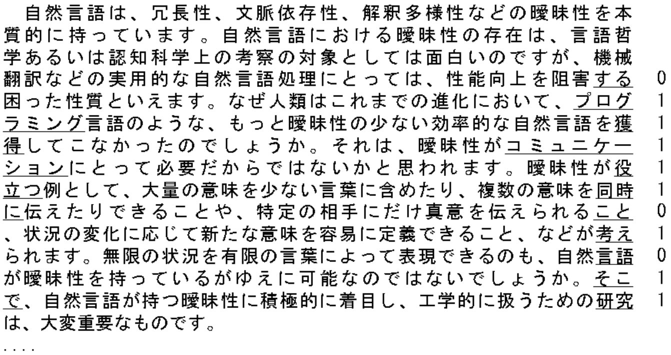
In this method the correspondence relationship between new-line position in each morpheme and the embedded bit is defined beforehand as the entry of a morphological-analysis dictionary, and secret data is embedded based on it. For example, if a word "suru" (do) is separated by a new-line code like "su|ru" ("|" denotes a new-line position), it will define corresponding to bit "1" of secret data. According to the standard line width specified at the stage of embedding procedure, secret data is embedded into morphemes that come near the end of the line. As shown in Fig. 5, long morphemes like "programming" and "communication" make two or more new-line positions correspond to bit "0" or bit "1", and it is made to start a new line with the number of characters which is not greatly different from the standard line width.



bit "0"	bit "1"	meaning in English
する↓	す↓る	do
プログラミング↓	プログラミン↓グ	programming
プロ↓グラミング	プログ↓ラミング	programming
獲得↓	獲↓得	obtain
コミュニケーション↓	コミュニケーショ↓ン	communication
コミュニケ↓ーション	コミュニケー↓ション	communication
コミュ↓ニケーション	コ↓ミュニケーション	communication
役立つ↓	役↓立つ	useful
と↓して	として↓	as
同時に↓	同時↓に	at the same time
こと↓	こ↓と	thing
考↓え	考え↓	opinion
言語↓	言↓語	language
そこで↓	そこ↓で	therefore
研↓究	研究↓	research

Fig. 5. Example of rule table corresponding to morphological headword and bit of secret data. (Each arrow denotes the place of new line code)

An example of stego text in which embedded information uses the corresponding table of Fig. 5 is shown in Fig. 6. Figure 6 is a justified document, and the variation in the number of characters of per one line can hardly be found visually. In the example of Fig. 6, the new-line positions of the first three lines are dummy in which secret data is not embedded, in the lines from the 4th to the 10th the flag sequence "0111110" is embedded, and at the 11th line the main part of embedded secret data "1011 ..." begins.



```
　自然言語は、冗長性、文脈依存性、解釈多様性などの曖昧性を本
質的に持っています。自然言語における曖昧性の存在は、言語哲
学あるいは認知科学上の考察の対象としては面白いのですが、機械
翻訳などの実用的な自然言語処理にとっては、性能向上を阻害する  0
困った性質といえます。なぜ人類はこれまでの進化において、プログ  1
ラミング言語のような、もっと曖昧性の少ない効率的な自然言語を獲  1
得してこなかったのでしょうか。それは、曖昧性がコミュニケー  1
ションにとって必要だからではないかと思われます。曖昧性が役  1
立つ例として、大量の意味を少ない言葉に含めたり、複数の意味を同時  1
に伝えたりできることや、特定の相手にだけ真意を伝えられること  0
、状況の変化に応じて新たな意味を容易に定義できること、などが考え  1
られます。無限の状況を有限の言葉によって表現できるのも、自然言語  0
が曖昧性を持っているがゆえに可能なのではないでしょうか。そこ  1
で、自然言語が持つ曖昧性に積極的に着目し、工学的に扱うための研究  1
は、大変重要なものです。
....
```

Fig. 6. Example of generated and justified stego text and embedded bits using the corresponding table shown in Fig. 5. (In order to show each morpheme shown in Fig. 5, the underlines are attached. The underlines are un-displaying in fact)

The method under this correspondence rule has the following features.

(1) Since how to Embedding per word can be defined, as compared with the method of Section 3.2.1, it is difficult to detect the rule of the correspondence relationship between the bit of embedding information, and a new line. Therefore, it is strong against an extraction attack.

(2) Since new-line position can be defined for every word, it is possible to carry out the definition that avoids a new line in unnatural position.

On the other hand, in order to achieve this technique, it is necessary to solve the problem of the extraction error caused by the error of morphological-analysis procedure, and the definition method in the case of a one-character morpheme.

4 Implemenation and Evaluation

The technique defined in Section 3.2.1 has been implemented on computer. The tool has been implemented using the Java language, which can run on various OS's. In this section the amount of embedded information is evaluated using the tool. In the proposed technique, bits of secret data are not embedded in the lines before the start flag, and the lines after the end flag. However, in this evaluation, the number of bits that can be embedded into all lines is defined as "the total amount of embedding".

In this technique, a corresponding table, standard line width, and minimum line width are specified as parameters on the occasion of embedding procedure. Moreover, on the occasion of extraction procedure, the same table and the same minimum line width as that used at the stage of embedding procedure are specified. In an agglutinative language, the position into which a new line is put is comparatively free. However, Japanese language has the procedure called "Japanese hyphenation" that avoids only a punctuation and a parenthesis being sent to the following line, and starting new line while being a number sequence is to avoid. If many restrictions about position of new line are defined, the document will be easier to read, namely, it will become a natural document. Instead, since the flexibility at the time of embedding secret data at the number of characters of each line narrows, the number of characters of each line varies greatly for every line, and the document will be an unnatural one. In order to compare this trade off, the following three methods have been implemented and evaluated according to the restrictions in the position of new line.

Method 1 – Priority is given to the homogeneity of line width.
In the method 1, except for restrictions in Japanese hyphenation, new lines are inserted so that each line of cover text may be in agreement with a standard line width as much as possible.

Method 2 – Inserting new lines in a specific type of character sequence are restricted.
The method 2 is a method using the restrictions that do not start a new line within a specific character sequence (numbers and alphabets) in addition to the restrictions in the method 1.

Method 3 – Priority is given to the boundary line of the same type of character sequence.
In Japanese there are three types of characters: Hiragana, Katakana, and Chinese characters. In addition to the restrictions in the method 2, the method 3 adds the restrictions that do not insert new lines into a character sequence of Chinese characters, Hiragana, or Katakana. Furthermore, if number of characters surrounded by parenthesises is 5 or less, new line code isn't inserted between the characters. Therefore, a great portion of new-line positions becomes the boundary line of character types (Chinese character/ hiragana/ katakana/ number/ alphabet). In Japanese, since the boundary of a character type e.g., between hiragana and a Chinese character, or between katakana and hiragana are the boundaries of clauses in many cases, and inserting new line in the boundary of clauses is desired from the viewpoint of the ease of reading.

In the corresponding table used for the evaluation, when line width is even, the bit of secret data is set to "1", and the bit is set to "0" when line width is odd. This is for measuring the amount of bits that can be embedded. In order to make an illegal decryption difficult, in an actual corresponding table the relationship between line width and a bit should be made random. The standard line width was set to 50 (equivalent to 25 double-byte characters), the minimum line width was set to 40, and the secret data was an 8-bit sequence "10110100." In this evaluation, the secret data was embedded repeatedly at all new lines excluding the line under the minimum line width from the beginning of a cover text.

The cover texts used for evaluation and each total rate of embedding, i.e., bandwidth, are shown in Table 1. Here, the total rate of embedding is defined as the value that divides the total number of embedded bits by 8 times of text size, which was converted into the value per bit.

According to the results of Table 1, the total number of embedding bits is proportional to the size of cover text mostly. Therefore, the total rate of embedding doesn't depend on the size of cover text, and it turns out that the total rate of embedding is almost fixed. Moreover, notably, the difference in the total rates of embedding by the types of text did not appear, and did not have most differences arising from the determination methods of a new-line position. Therefore, we can conclude to be the proposed technique whose amount of embeddable information is stable.

Table 1. Cover texts used for the evaluation and each total embedding rate

Type of Text		Size of Text (byte)	Genre or Title	the total number of embedded bits (unit is bit)			total rate of embedding(5)		
				Method 1	Method 2	Method 3	Method 1	Method 2	Method 3
News	General	1,929		36	37	34	0.23	0.24	0.22
		1,751		33	33	32	0.24	0.24	0.23
	Special field	2,258	News article about cryptology	39	40	38	0.22	0.22	0.21
		2,433	News article about MS-Windows	45	45	44	0.23	0.23	0.23
	For Kids	3,765	News commentaryfor Kids	68	68	58	0.23	0.23	0.19
Technical Paper	Special field	2,290	Technical Paper in Japanese Conference	40	40	39	0.22	0.22	0.21
		3,336	Technical Paper in Japanese Conference	62	66	62	0.23	0.25	0.23
Novel	Classical	3,789	“Makurano soshi”	74	80	70	0.24	0.26	0.23
		6,353	“Genji Monogatari”	121	123	114	0.24	0.24	0.22
	For Kids	3,606	“Alice's Adventure in Wonderland" (Japanese Edition)	72	73	59	0.25	0.25	0.20
		5,418	“Kaze no Matasaburo”	105	106	89	0.24	0.24	0.21
	General	5,640	“Wagahai wa neko dearu”	110	109	106	0.24	0.24	0.23
		1,866	“Rashoumon Gate”	36	39	34	0.24	0.26	0.23

5 Discussion and Conclusion

Differences in tampering with documents cannot be found between cover text and stego text when it is assumed that an attacker is able to get and compare both cover text and stego text, or to get and compare two or more stego texts. When it is assumed that attacker can get only a single stego text, the stego text needs to be as natural as possible. The proposed technique is effective in case that attacker can get only a single stego text.

In the proposed technique, if a correction that changes the position of new lines is made to a document, secret data will be eliminated. The threat of elimination is unavoidable, because to maintain naturalness so that the embedding will be hard to detect, the secret data has to be embedded only in new line codes. We consider that making it hard to detect the embedding is effective in reducing the threat of intentional elimination. The mechanical correction that arranges the number of characters of each line is made by mailer in many cases, the ends of the lines are only turned up and the original new-line positions are saved. In the proposed technique, the correction which deletes original new-line codes or regives new-line positions serves as a threat. Since the mechanical distinction with the text and titles of chapters or itemized statements is difficult, this correction is seldom made in application software for plain text. Therefore, a possibility that new-line positions may be changed by mechanical correction is considered to be infrequent. In general, attackers cannot tamper with hard copy. In the proposed technique, secret data remains also in hard copy. Moreover, if secret data is encrypted and embedded, the tolerance to extraction attack can be increased.

In the proposed technique, the message sender and recipient must share the same secret rule table. The characteristic is inconvenient especially for n-to-n communication. However in case that the technique is applied to fingerprinting, the problem doesn't occur, because recipient is the verifier, i.e., the sender.

As a future work, the total rate of embedding should be increased.

References

1. J.T.Brassil, S.Low, N.F.Maxemchuk, L.O'Gorman, “Electronic Marking and Identification Techniques to Discourage Document Copying,” Proc. IEEE INFOCOM '94, Vol.3, pp.1278-1287.
2. M.Kwan, “The SNOW Home Page,” See http://www.darkside.com.au/snow/.
3. T.Matsumoto, H.Itoyama, “Can Bypassing Lawful Access be Always Detected?,” Technical Report of IEICE, ISEC96-79, pp. 159-164, Mar. 1997 (in Japanese).
4. T.Matsumoto, H.Nakagawa, I.Murase, “Information hiding technical development for network-development of finger printing system for document -FinPri.txt,” Information-Technology Promotion Agency, Jun. 2000 (in Japanese).
5. M.Chapman, G.Davida, “Hiding the Hidden: A Software System for Concealing Ciphertext as Innocuous Text,” Proc. Int. Conf on Information and Communicatons Security, LNCS 1334, pp.335-345, Springer, 1997.
6. K.Maher, “Texto,” See http://www.eberl.net/textodemo.html.
7. M.J.Atallah, V.Raskin, M.Crogan, C.Hempelmann, F.Kerschbaum, D.Mohamed, S.Naik, “Natural Language Watermarking: Design, Analysis, and a Proof-of-Concept Implementation,” Proc Int. Workshop IH 2001, LNCS 2137, pp.185-199, Springer, 2001.
8. M.J.Atallah, V.Raskin, C.F.Hempelmann, M.Karahan, R.Sion, U.Topkara, K.E.Trizenberg, “Natural Language Watermarking and Tamperproofing,” Proc. Int. Workshop IH 2002, LNCS 2578, pp.196-212, Springer, 2002.
9. Igor A. Bolshakov, “A Method of Linguistic Steganography Based on Collocationally-Verified Synonymy,” Proc. Int. Workshop IH 2004, LNCS 3200, pp.180-191, Springer, 2004.