証明可能ステガノグラフィを用いたチャットへの情報埋め込み実験

An Experiment on Embedding Information into Chat Messages Using Provably Secure Steganography

梅澤 克之 [1]　ウォルゲムト スベン　古川 一夫 [2]　寶木 和夫 [3]
Katsuyuki Umezawa　Sven Wohlgemuth　Kazuo Furukawa　Kazuo Takaragi

湘南工科大学 [1]　関西大学 [2]　(株) ハイセーフ [3]
Shonan Institute of Tech.　Kansai University　HISAFE, Ltd.

1　まえがき

近年「完全安全ステガノグラフィ」と呼ぶ技術が提案された [1]．この技術では，情報を埋め込む前の文書（カバーテキスト）と埋め込み後の文書（ステゴテキスト）の統計的性質の分布が完全に一致する（KL ダイバージェンスがゼロになる）ことを実現している．これにより情報が埋め込まれたこと自体を秘匿される．さらに，疎サンプリングに基づく効率的なステガノグラフィ（SparSamp）が提案された [2]．この方法も元の確率分布を保持するため安全性を保障した上で計算量を低減した方法である．SparSamp は内部で使用する生成モデルをプラグ・アンド・プレイ設計になっており様々な生成モデルを置き換えることが可能であるという特徴を持つ．生成 AI が登場することにより，無限のサンプリングが可能となったことで，このような様々なステガノグラフィ技術が提案され始めた．しかし，このような技術を用いても，方言や老け台詞などの特徴を持つ文書に対して，人が読んだ時に違和感があり情報が埋め込まれていることがばれてしまう可能性がある．また，今後電子メールなどが廃れ代替としてチャットや SNS などがコミュニケーションの主要な手段になる世界では，極短い文書に情報を埋め込む必要が生じることが想定できる．本研究の目的は，チャットでのメッセージ交換のような極短い文章に対して SparSamp を適用した際に，チャットとしての違和感を抱くか否かを実験によって確かめることである．

2　従来研究

2.1　証明可能な安全性を備えたステガノグラフィ（Provably Secure Steganography: PSS）

算術符号化（Arithmetic Coding: AC）ベースのステガノグラフィーとしていくつかの方式が提案されている（例えば [3]）．これらの手法は長い列に対して特に有効であり情報エントロピー値に近い圧縮率を達成できる．Kaptchuk ら [4] は，AC ベース方式における「ランダム性の再利用問題」と呼ばれる課題を指摘し「Meteor」を提案した．また，Witt ら [1] は最小エントロピー結合の概念を用いた情報理論的ステガノグラフィーを提案し，完全なステガノグラフィの安全性が結合問題に等価であること，完全に安全なシステムにおける最大伝送効率の達成が最小エントロピー結合問題に等価であることを証明した．

2.2　SparSamp

Wang ら [2] によって，疎サンプリングに基づく効率的なステガノグラフィ（SparSamp）が提案された．この方法は元の確率分布を保持することで安全性を保証しつつ，計算量を低減するものである．SparSamp におけるメッセージビット（111）の埋め込みの例を図 1 に示す．

まず，最初のステップで，疑似乱数生成器より $r_i = 0.575$ を得る．$r_i(111) = (r_i + 7/8) mod 1 = 0.450$，$r_i(110) = (r_i + 6/8) mod 1 = 0.325$，$r_i(000) = (r_i + 0/8) mod 1 = 0.575$ より，“Hello” が選択されるが，候補が “110”，“111”，“000” の 3 つとなりこのままだと復号できない．よって次のステップに進む．候補が 3 つに絞られ，3 つの中で “111” のインデックスは 1 なので，$r_{i+1}(111) = (r_i + 1/3) mod 1 = 0.087$ となり，“World” が選択される．送信者と受信者は同じ生成モデル，カバーテキスト，擬似乱数生成器，および鍵を保持する必要がある．送信者はメッセージ全体が埋め込まれるまでメッセージと擬似乱数を用いて次のトークンを継続的にサンプリングする．ステゴテキストが作成されると受信者に送信される．受信者は送信者の状態と同期し，逆のプロセスを用いてメッセージを抽出できる．受信者は埋め込まれるビット数を知っていればステゴテキストの終端記号は不要である．

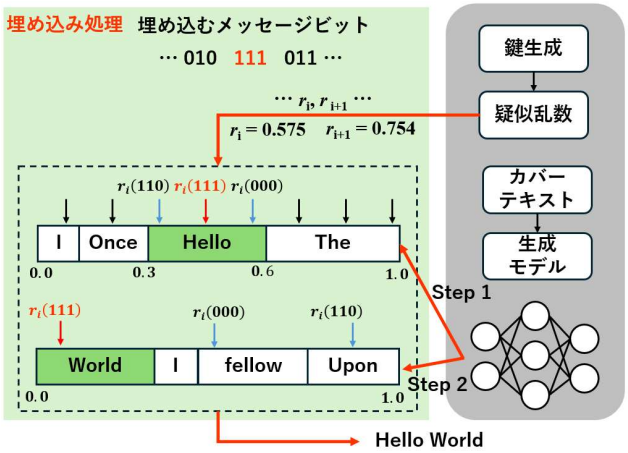

図 1　SparSamp におけるメッセージビットの埋め込み（従来研究 [2] から著者が作成）

3　提案手法

チャットアプリのような短い文章をやり取りする状況において，その短い文章に情報を埋め込むことを提案す

る（図 2 参照）．まず情報の受信者のいくつかのメッセージ（2〜4 つ）とそれに反応する情報の送信者の 1 つのメッセージをまとめてカバーテキスト（情報を埋め込む前のテキスト）とする．その時各メッセージには，適当な区切り文字（例えば「。」を連接する）．そのカバーテキストをもとに 4 ビットの情報を埋め込んだステゴテキストを生成する．生成されたステゴテキストは数文字であり，かつ，文章や単語の途中で中途半端に切られている．そこで，情報の送信者は、ステゴテキストに続けて自然な文章になるように，文書を追加する。それに引き続き本来送信したいメッセージを送信する．ここまでで 1 つのステップ（4 ビットの情報の送信）が完了する．これを繰り返すことで 4 ビットすつ情報を秘密裏に送信可能となる。

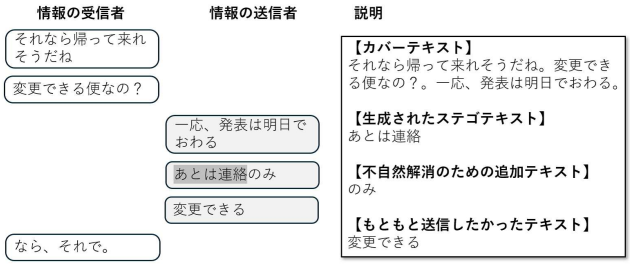

図 2　チャットにおけるステゴテキスト生成

4　実験方法

チャットアプリを使った実際の会話をもとに，3 節で説明した方法で 4 ビットの情報を埋め込んだステゴテキスト 14 個生成した（4 × 14=56 ビット）．全メッセージは 77 個（通常メッセージが 63 個，ステゴメッセージが 14 個）である．通常メッセージとステゴメッセージが区別がつかないようにして，被験者に違和感の調査を行った．条件として，2 名のうち片方は日本に居て，もう片方はタイにいる，という情報だけ与えた．その後，(実験 1) チャットでの会話としてどこかに違和感はないか，(実験 2) どこかに情報が埋め込まれているということだけを伝えた上で違和感はないか，の 2 つの質問を行い，その個所を丸印で囲ってもらった．実験参加者は普段からチャットアプリは良く使う 19 名である．

5　実験結果

今回は再現率（Recall：TP/(TP+FN)）と偽陰性率（FNR: FN/(FN+TP)）で評価する．ここで TP は真陽性（ステゴメッセージをステゴメッセージと正しく判断できた個数），FN は偽陰性（陽性を誤って陰性と判断，つまりステゴメッセージを通常メッセージと判断した個数）である．再現率は，実際に陽性であるもののうち正しく陽性と予測できた割合，つまり，ステゴメッセージを正しく指摘できた割合である．また，偽陰性率は，実際に陽性なのに誤って陰性とした割合，つまりステゴメッセージの見逃しの多さを表す．表 1 より，今回の実験では，再現率が低く、偽陰性率が高かった．これにより，ステゴテキストの箇所を指摘しにくい，つまり違和感を抱かない，ということが確認できた．さらに情報が埋め込まれているということを伝えたうえでの実験（実験 2）であっても，実験 1 とほとんど変わらない再現率と偽陰性率であった．

表 1　実験結果（19 人の平均値）

	実験 1	実験 2
指摘箇所数	9.7	11.8
正解数	3.9	4.0
不正解数	5.8	7.8
TP（真陽性）	3.9	4.0
FP（偽陽性）	5.8	7.8
FN（偽陰性）	10.1	10.0
TN（真陰性）	57.2	55.2
Accuracy（正解率）	0.793	0.768
Precision（適合率）	0.434	0.411
Recall（再現率）	0.278	0.286
F 値（F1-score）	0.311	0.312
TNR（真陰性率）	0.907	0.876
FPR（偽陽性率）	0.093	0.124
FNR（偽陰性率）	0.722	0.714

6　まとめと今後の課題

本研究では，短文チャットメッセージに情報を埋め込み，日本語として違和感を抱くか否かを実験を通して検証した．結果として再現率が低く偽陰性率が高くなり，つまり情報が隠されているステゴテキストでも違和感を抱きにくいことが分かった．

短文チャットへの SparSamp 適用は，人間の直観による検出を回避し得る可能性を示したが，統計的検証の強化（被験者増，対照群の拡充），自動検出器との比較，言語多様性の検討，倫理的配慮（悪用防止）等が今後の課題である．

参考文献

[1] Christian Schroeder de Witt, Samuel Sokata, J. Zico Kolter, Jakob Forester, and Martin Strohmeier, “Perfectly Secure Steganography Using Minimum Entropy Coupling,” Published as a conference paper at ICLR 2023.

[2] Yaofei Wang, Gang Pei, Kejiang Chen, Jinyang Ding, Chao Pan, Weilong Pang, Donghui Hu, and Weiming Zhang, “SparSamp: Efficient Provably Secure Steganography Based on Sparse Sampling,” Proceedings of the 34th USENIX Conference on Security Symposium, USENIX Association, p.p. 1–19, 2025.

[3] R.J. Anderson and F.A.P. Petitcolas, “On the limits of steganography,” IEEE Journal on Selected Areas in Communications, 16(4), p.p. 474-?481, 1998.

[4] Gabriel Kaptchuk, Tushar M. Jois, Matthew Green, and Aviel D. Rubin, “Meteor: Cryptographically secure steganography for realistic distributions,” Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021.