Detection of substitution-based linguistic steganography by relative frequency analysis

Zhili Chen\textsuperscript{a,b,*}, Liusheng Huang\textsuperscript{a,b}, Wei Yang\textsuperscript{a,b}

\textsuperscript{a}NHPCC, School of CS. & Tech., USTC, Hefei 230027, China
\textsuperscript{b}Suzhou Institute for Advanced Study, USTC, Suzhou 215123, China

\textbf{Article Info}

Article history:
Received 5 August 2010
Received in revised form
17 January 2011
Accepted 16 March 2011

Keywords:
Information hiding
Linguistic steganography
Relative frequency analysis
Detection
Substitution-based
Synonym-substitution

\textbf{Abstract}

Linguistic steganography hides information in natural language texts. Because of the increasing in importance and quantity of natural language texts, linguistic steganography plays a more and more important role in Information Security (IS) area today. Substitution-based linguistic steganography is one of the most commonly used linguistic steganography methods, which is of considerable security and favorable simpleness. In this paper, we propose a straightforward method based on Relative Frequency Analysis (RFA), which makes use of the frequency characteristics of the testing texts (the texts being tested), to detect substitution-based linguistic steganography. We formally prove several properties about relative frequency which can be used in the detection process and propose a detection scheme. And then as an example, an existent synonym-substitution system T-Lex is examined and the detection experiment is carried out. In the experiment with pure literature texts, the accuracy, precision and recall of the detection are found to be as high as 98.64%, 97.77% and 99.55%, respectively, when the substitution count is 90, while in the experiment with balanced texts, the highest detection accuracy is 95%, which indicates that the detection scheme is promising.

\textcopyright 2011 Elsevier Ltd. All rights reserved.

1. Introduction

Nowadays, natural language texts have become one of the largest chunks of digital data that people encounter daily. Numberless articles in digital format come from many newspapers, magazines, scientific journals and conferences every year. Besides, emails, blogs and all kinds of web pages online provide even more textual data. This enhancement in the significance and quantity of digital text on Internet creates increased concerns about the usage of textual data as a covert channel of communication. One of such covert ways of communication is known as linguistic steganography. Linguistic steganography embeds messages into natural language texts in a covert manner such that the presence of the hidden messages cannot be easily discovered by anyone except the intended recipient.

Linguistic steganography methods can be grouped under two categories (Bennett, 2004), Linguistic Driven Generation-based (LDG-based) and Linguistic Driven Modification-based (LDM-based). The first group is based on generating a new stego-text for a given message. For example, NICETEXT (Chapman, 1997; Chapman et al., 2001) and TEXTO (Maher, 1995) are LDG-based. The second group is based on modification of an existent covert text. Coarsely speaking, LDG-based methods, which generate stego-texts looking like natural texts but being lack of coherent sense, have a higher hiding capacity than LDM-based methods. LDG-based methods are mainly used to conceal a great bulk of encrypted...
secret information during data transmission. LDM-based methods have a lower hiding capacity and their stego-texts look like natural texts both syntactically and semantically. They are used for hiding more secret information transferred and are widely used in natural language watermarking. One subcategory of LDM-based linguistic steganography is called substitution-based linguistic steganography which substitutes some elements of the cover text with their semantically equivalent ones. In this paper, we mainly focus on the detection of substitution-based linguistic steganography.

This paper presents a method making use of relative frequency analysis to detect substitution-based linguistic steganography. Different from the previous detections (Taskiran et al., 2006; Luo et al., 2008; Yu et al., 2008) which usually need the context information, we consider the detection utilizing frequency information only. We closely look into the frequency characteristics of the substitution elements in the substitution sets in both cases of normal texts and stego-texts and find that the relative frequency satisfies certain properties. We then propose a detection scheme making use of these properties. Finally, we examine the synonym-substitution system implemented by Weinstein (1999) as an example, design a detection algorithm under the detection scheme and validate the detection by some experiments. Experimental results show the efficiency of the detection.

2. Related work

Substitution-based linguistic steganography can be classified into synonym substitution-based methods (Weinstein, 1999; Atallah et al., 2000; Bergmair, 2004; Bolshakov and Gelbukh, 2004; Bolshakov, 2004; Calvo and Bolshakov, 2004; Topkara et al., 2006; Liu et al., 2007), semantically equivalent rule substitution-based methods (Hugg, 1999), synonymous sentence substitution-based methods (Murphy, 2001), translation-based methods (Grothoff et al., 2005a, b; Stutsman et al., 2006) and so on, according to the substitution element. Among those, synonym substitution-based linguistic steganography is most widely used. In synonym substitution system, the hidden message is embedded by substituting a word with one of its synonyms. The stego-text keeps the same sense before and after substitution. In this section, we introduce T-Lex system as an example of substitution-based linguistic steganography and discuss its drawbacks that make the accurate detection possible. After this, we analyze the previous attacks and discuss their weakness.

2.1. T-Lex system

The most important problem that synonym substitution faces is how to define the synonym set. In natural language, words often have many senses in different contexts. How to determine the exact sense in a certain context is a hard problem known as word sense disambiguation in Natural Language Processing (NLP). The definition of synonym set must guarantee that all synonym sets are mutually disjoint in order not to cause word sense disambiguation problem. However, the multi-sense property makes definition of synonym set difficult. For example, word A and B are synonyms in a context, word B and C are also synonyms in another context, but word A and C can have different senses in any context.

Winstein (1999) proposed a solution for synonym set definition in T-Lex system. He used WordNet (WordNet) to select synonyms with correct senses. In WordNet, Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synset), each of which expressing a distinct concept. Synsets are interlinked both conceptually-semantically and lexically. In T-Lex system, not all synonyms in WordNet database are included in the synonym database. Only those words completely in the same synsets are grouped in the same synonym set. For example, assume that words a, b, c only belong to the synsets S1={a, b}, S2={a, b, c}. In this case, even though both words a and b have more than one sense, they still can be interchanged semantically in all contexts. Applying the criteria described above, Keith obtained synsets containing about 30% of 70,803 single word entries in WordNet as the synonym database of T-Lex system. The average synset size is 2.56 while the maximum is 13 and the size is 2 (Winstein, 1999).

T-Lex system currently only hides text messages in the cover texts, but modification of hiding any kind of messages is easy. A given text message is embedded into the cover text using the synset database as follows. First, the letters of the message text are Huffman coded according to English letter frequencies. Then, the Huffman code binary string is represented in mixed-base form. As a simple example, suppose that the binary string to be embedded is (010)2 and that the message text are Huffman coded according to English letter frequencies. Then, the Huffman code binary string is represented in mixed-base form, each digit has a different base. For (010)2, a context, word A and B are synonyms in a context, word B and C are also synonyms in another context, but word A and C can have different senses in any context.

Winstein (1999) proposed a solution for synonym set definition in T-Lex system. He used WordNet (WordNet) to select synonyms with correct senses. In WordNet, Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synset), each of which expressing a distinct concept. Synsets are interlinked both conceptually-semantically and lexically. In T-Lex system, not all synonyms in WordNet database are included in the synonym database. Only those words completely in the same synsets are grouped in the same synonym set. For example, assume that words a, b, c only belong to the synsets S1={a, b}, S2={a, b, c}. In this case, even though both words a and b have more than one sense, they still can be interchanged semantically in all contexts. Applying the criteria described above, Keith obtained synsets containing about 30% of 70,803 single word entries in WordNet as the synonym database of T-Lex system. The average synset size is 2.56 while the maximum is 13 and the size is 2 (Winstein, 1999).

T-Lex system currently only hides text messages in the cover texts, but modification of hiding any kind of messages is easy. A given text message is embedded into the cover text using the synset database as follows. First, the letters of the message text are Huffman coded according to English letter frequencies. Then, the Huffman code binary string is represented in mixed-base form. As a simple example, suppose that the binary string to be embedded is (010)2, and that currently the sentences following are being considered.

"... A bicycle was lying upon the \{roadside, grass\} ..."

"... and he had a pair of \{shrewdly, astutely, sagaciously\}, careless boyish eves..."

In the example, the first words with no number leading in the brace are the original words to be replaced and the numbered words constitute their corresponding synonym set. In mixed-base form, each digit has a different base. For (010)2 = 2, we have

\[ 4a_1 + a_0 = 2 \]

With the constraints \( 0 \leq a_0 < 4 \) and \( 0 \leq a_1 < 2 \). Thus, we get \( a_0 = 2 \) and \( a_1 = 0 \). This indicates that "roadside" and "shrewdly" should be replaced by "wayside" and "sagaciously".

2.2. Drawbacks of T-Lex system

Two shortcomings of T-Lex system have been pointed out (Taskiran et al., 2006). One is that it sometimes substitutes words with their synonyms that do not agree with the correct English usage; the other is that the words after substitution do not agree with the genre and the author style of the cover text.
Another shortcoming of T-Lex system is that it often substitutes words with their synonyms whose word frequencies are far different from the original word frequencies. For instance, in the example shown above the synonym sets with word frequencies are

\{wayside: 206, roadside: 242\}
\{shrewdly: 119, astutely: 8, sagaciously: 20, sapiently: 1\}

The substitutions by T-Lex system, that is “roadside” and “shrewdly” are replaced by “wayside” and “sagaciously”, cause the word frequencies to change from 242 and 119 to 206 and 20 respectively. As we will see, our detection mainly exploits the last shortcoming.

2.3. **Previous detection methods of T-Lex system**

Though the hiding principle of T-Lex is simple and straightforward, the detection of T-Lex is really challenging. There are some reasons for this. The most important one is that T-Lex system modifies few synonym words out of the cover texts. For example, for a cover text of thousands of words there may be only tens of synonym words that can be used for substitution and tens of bits can be hidden in it. As a result, the stego text is nearly the same as the cover text and the distinguishing of them is hard. Another reason is that the substitutions of synonym words usually keep the syntax and semantics correctly and it is challenging to determine whether the synonym words in the testing text is replaced by another one or not.

Although it is hard for the detection of T-Lex system, some detection methods of T-Lex system have been proposed, as far as we know. They are the detections based on language model and support vector machine (DLM-SVM) (Taskiran et al., 2006), based on synonym pair (DSP) (Luo et al., 2008) and based on context information (DCI) (Yu et al., 2008).

The DLM-SVM-based detection utilizes the language model to obtain some classification features of the testing sentences, and then uses SVM classifier to classify them to normal sentences and stego sentences. DLM-SVM detects the application of T-Lex system without using the synonym dictionary, that is, the detection is blind, which is pretty difficult. However, the DLM-SVM based detection detects the sentences instead of texts, which more or less makes the detection easier than the detection of whole text. It is reported that “the accuracy on steganographically modified sentences was found to be 84.9% and that for unmodified sentences to be 38.6%” (Taskiran et al., 2006). That is to say DLM-SVM has a high recall of 84.9%, but has a poor accuracy and precision.

The DSP-based detection introduces the notion of synonym pair and uses it to detect the Chinese texts which may be processed for hiding information utilizing synonym substitution algorithm like T-Lex system. In this detection, the synonym dictionary used by the hiding algorithm is necessary. The experiment shows that the false negative rate is approximately 4% and the false positive rate is approximately 9.8%, so the accuracy is about 86.2% (Luo et al., 2008).

The DCI-based detection uses the context information of synonyms obtained by Google search engine to obtain the present probability of each synonym in the synonym set and then use the expected value of the synonyms in the testing text to detect the stego texts processed by T-Lex system. Also, the synonym dictionary is needed for the detection. Experiment shows that the accuracy is about 90% (Yu et al., 2008).

As we can see, although some detection methods have been proposed, the detection performance still needs to be enhanced even when the synonym dictionary is used. The detection proposed in the paper can greatly improve the detection performance when the synonym dictionary is available.

3. **Theoretical analysis**

In order to present the analysis clearly, we need to firstly introduce some definitions and the notation in the sections 3.1 and 3.2.

3.1. **Definitions**

3.1.1. **Substitution element**

Certain part of natural language that is suitable for substitution with its equivalent ones. A substitution element may be a word, a phrase, a sentence and so on, depending on the substitution-based linguistic steganography method used, e.g., in the scenario of synonym substitution, a substitution element should be a synonym word.

3.1.2. **Substitution set**

The set of substitution elements that can be exchanged each other. Substitution elements are grouped into sets so that the elements in the same set are equivalent in certain aspects, such as syntax, semantics and the like. In the case of synonym substitution, the synonym set is the substitution set.

3.1.3. **Substitution dictionary (SD)**

The dictionary containing all the available substitution sets, e.g., in case of synonym substitution, the substitution dictionary should be the synonym dictionary.

3.1.4. **Substitution list**

A list of substitution elements from the SD used in a certain text when applying substitution-based linguistic steganography methods, e.g., in the case of synonym substitution, the synonym words such as “unwedded”, “seamanly” and “gladdened” are found in a text in all, then the synonym sets [unwedded, unwed], [seamanly, seamanlike] and [gladdened, exhilarated] make up the substitution list of the text.

3.1.5. **Natural frequency**

The frequency of a substitution element in the normal texts, normally obtained by processing a large corpus, e.g., if we use BNC corpus in our experiments, the frequency of the synonym word “unwedded” in BNC is 2, then its natural frequency is also 2.

3.1.6. **Substitution frequency**

The frequency of a substitution element in the stego texts, the distribution of which can be inferred from the substitution-based linguistic steganography method used, e.g., in the
synonym substitution steganography, if the secret bit string is completely random, the elements in the same substitution set are equally selected for substitution, and then we can infer that the substitution frequencies of the elements in the same substitution set in the stego texts are all equal. It also can be obtained by processing a large number of stego texts, just as the obtaining of natural frequency.

3.1.7. Relative frequency
The relative occurrence frequency of each substitution element within a certain substitution set, which is evaluated by computing the proportion of its natural frequency in case of normal text (called normal relative frequency) or its substitution frequency in the case of stego text (called substitution relative frequency), e.g., in the synonym set with natural frequencies (timeworn: 5, hackneyed: 45, trite: 92), the natural relative frequencies of “timeworn”, “hackneyed” and “trite” are 0.035, 0.317 and 0.648, respectively.

3.1.8. Natural relative frequency (NRF) score
A score of a substitution element in direct proportion to its NRF value, e.g., for a substitution set with NRF values as [abutting: 0.03, adjoining: 0.72, bordering: 0.25], the scores of “abutting”, “adjoining” and “bordering” are 3, 72 and 25, respectively, if the proportion coefficient is 100.

3.1.9. NRF unbalance
The NRF unbalance of a substitution set is the condition on which the NRFs of the elements in the same substitution set are not all equal. The more the differences between different NRFs are, the more NRF unbalanced the substitution set is. E.g., the substitution set with NRFs [preconception: 0.88, prepossession: 0.12] is NRF unbalanced, while the set [ult: 0.5, ultimo: 0.5] is NRF balanced, the set [untie: 0.89, unlace: 0.11] is more NRF unbalanced than the set [preconception: 0.88, prepossession: 0.12], while the set [rollickingly: 0, boisterously: 1] is the most NRF unbalanced.

The NRF unbalance of a text refers to the condition on which at least one of the substitution sets in the substitution list of the text is of NRF unbalance. The more NRF unbalanced the substitution sets in the substitution list of a text are, the more NRF unbalanced of the text is.

3.2. Notation

$T^{(N)}$: A normal text which has not been embedded information using substitution-based linguistic steganography.

$T^{(S)}$: A stego-text which is obtained by embedding information into a normal text using substitution-based linguistic steganography.

$D$: The substitution dictionary used by substitution-based linguistic steganography.

$L$: The substitution list, $L = \{S_0, S_1, \ldots, S_{n-1}\}$ where $n$ is the count of substitutions.

$S_i$: The substitution set, $S_i = \{s_{i0}, s_{i1}, \ldots, s_{im_{i-1}}\}$ ($0 \leq i < n$) where $s_{ij}$ ($0 \leq j < m_i$) denotes a substitution element and $m_i$ represents the size of substitution set.

$F_i$: $S_i$’s corresponding natural frequency set, denoted by $F_i = \{f_{i0}, f_{i1}, \ldots, f_{im_{i-1}}\}$.

$P_i$: The NRF set of substitution set $S_i$ denoted by $P_i = \{p_{i0}, p_{i1}, \ldots, p_{im_{i-1}}\}$.

$Q_i$: The substitution relative frequency set of substitution set $S_i$, denoted by $Q_i = \{q_{i0}, q_{i1}, \ldots, q_{im_{i-1}}\}$.

$C_i$: The NRF score of substitution set $S_i$ denoted by $C_i = \{c_{i0}, c_{i1}, \ldots, c_{im_{i-1}}\}$.

$E_{Si}^{(N)}$: The NRF score expected value of $S_i$ in case of normal texts.

$E_{Si}^{(S)}$: The NRF score expected value of $S_i$ in case of stego-texts.

$V_{Si}^{(N)}$: The NRF score variance of the normal text $T^{(N)}$.

$V_{Si}^{(S)}$: The NRF score variance of the stego-text $T^{(S)}$.

3.3. Detection task

We consider a testing text $T$ probably with secret information hidden using substitution-based linguistic steganography. Assume that a possible substitution dictionary $D$ has already been obtained anyway and the hidden information is a random binary string. The assumption is reasonable, as one can guess the substitution dictionary given the testing texts and usually the hidden information is encrypted or compressed data which can be regarded as random binary string. The detection task is to judge whether the text $T$ is a stego-text using the substitution dictionary $D$ and the random binary string assumption together with other available resources.

3.4. Relative frequency analysis

Now, it is time for us to begin the analysis. We first consider the NRF score expected value of a substitution set, in the cases of both normal texts and stego-texts. We find that the expected value in the case of normal texts tends to be greater. We then infer that the NRF score expected value of a normal text also tends to be greater than that of a stego-text and find that as the NRF unbalance of the substitution sets gets stronger, the difference becomes even greater. Finally, we discover that the NRF score variance of a stego-text appears to be statistically greater than that of a normal text and the difference gets greater as the NRF unbalance enhances.

First of all, we examine the NRF score expected values of a substitution set in cases of stego-texts and normal texts. Given the substitution set is $S_k = \{s_{k0}, s_{k1}, \ldots, s_{km_{-1}}\}$ and the natural frequency set corresponding is $F_k = \{f_{k0}, f_{k1}, \ldots, f_{km_{-1}}\}$, where $k < n$ is the number of substitution set in its substitution list, $m_i$ is the count of the substitution elements in the set, the NRF of each substitution element $s_{ki}$ is

$$p_{ki} = \frac{f_{ki}}{\sum_{j=0}^{m_{-1}} f_{kj}} \quad 0 \leq i < m_k \quad (1)$$

Obviously, $\sum_{i=0}^{m_{-1}} p_{ki} = 1$ holds.

As we expect to find the frequency characteristics among different substitution sets, it is necessary to assign a reasonable value to each of the substitution element indicating its frequency property. For this purpose, we formulate the NRF score of each substitution element $s_{ki}$ as follows.

$$c_{ki} = g(p_{ki}) \quad 0 \leq i < m_k \quad (2)$$
where function $g(a)$ is a strictly increasing function, that is to say, for any integers $a, b (0 \leq a < b < m_k)$, $g(p_k) < g(p_h)$ if and only if $p_k < p_h$. In this paper, we simply define $g(p_k) = \kappa \cdot p_k$, $\kappa > 0$, so
\[ c_k = \kappa \cdot p_k, \quad 0 \leq i < m_k, \quad \kappa > 0 \]  
(3)

We are then able to calculate the NRF score expected value of $S_k$ in case of normal texts.

\[ E_{k}^{(N)} = \frac{1}{n} \sum_{i=0}^{m_k-1} p_k \cdot c_k = \kappa \cdot \frac{1}{m_k} \sum_{i=0}^{m_k-1} p_k = \frac{\kappa}{m_k} \quad 0 \leq i < m_k \]  
(4)

In the stego-texts, secret information is usually encoded by simply selecting which substitution element to appear disregarding the NRFs of the substitution set. According to the random binary string assumption in Section 2.3, we can infer that the substitution elements in a substitution set are selected in equal possibility. As a result, the substitution relative frequency of each substitution element $s_{ki}$ in the substitution set $S_k$ is

\[ q_{ki} = \frac{1}{m_k}, \quad 0 \leq i < m_k \]  
(5)

And then the NRF score expected value of the substitution set $S_k$ in the case of stego-texts is

\[ E_{k}^{(S)} = \frac{1}{n} \sum_{i=0}^{m_k-1} q_{ki} \cdot c_{ki} = \kappa \cdot \frac{1}{m_k} \sum_{i=0}^{m_k-1} p_k = \frac{\kappa}{m_k} \quad 0 \leq i < m_k \]  
(6)

Theorem 1. If substitution set $S_k$ is of NRF unbalance, then $E_{k}^{(N)} > E_{k}^{(S)}$.

Proof. Given the substitution set $S_k = \{s_{k0}, s_{k1}, \ldots, s_{km_k-1}\} (0 \leq k < n)$, and its corresponding NRF sets $p_{k0} = p_{k1} = \ldots = p_{km_k-1}$ which satisfy $p_{k0} < p_{k1} < \ldots < p_{km_k-1}$, as the substitution set is of NRF unbalance, at least one equal sign in the satisfying inequality does not hold. Assume that $p_{k0} < p_{k1}$ without loss of generality, according to equation (3), we have $c_{k0} < c_{k1} < \ldots < c_{km_k-1}$, then

\[ E_{k}^{(N)} - E_{k}^{(S)} = \frac{1}{m_k} \sum_{i=0}^{m_k-1} p_k \cdot c_{ki} - \frac{1}{m_k} \sum_{i=0}^{m_k-1} q_{ki} \cdot c_{ki} = \frac{1}{m_k} \sum_{i=0}^{m_k-1} (p_k - q_k) \cdot c_{ki} \]

As $\sum_{i=0}^{m_k-1} p_k = 1$, $q_k = 1/m_k$ and $p_{k0} < p_{k1}$ holds, we can find $p_{k0} < h < m_k$ satisfying $p_{kh} > \frac{1}{m_k} = q_k$ and $p_{kh-1} < 1/m_k$. So

\[ E_{k}^{(N)} - E_{k}^{(S)} = \frac{1}{m_k} \sum_{i=0}^{h-1} (p_k - q_k) \cdot c_{ki} + \frac{1}{m_k} \sum_{i=h}^{m_k-1} (p_k - q_k) \cdot c_{ki} \]
\[ > \frac{1}{m_k} \sum_{i=0}^{h-1} (p_k - q_k) \cdot c_{ki} + \frac{1}{m_k} \sum_{i=h}^{m_k-1} (p_k - q_k) \cdot c_{ki} \]
\[ = \frac{1}{m_k} \sum_{i=0}^{h-1} (p_k - q_k) \cdot c_{ki} + \frac{1}{m_k} \sum_{i=h}^{m_k-1} (p_k - q_k) \cdot c_{ki} = 0 \]

Thus $E_{k}^{(N)} > E_{k}^{(S)}$.

Corollary 1. A normal text has a greater NRF score expected value than a stego-text if they are of NRF unbalance and have the same substitution list, namely $E_{k}^{(N)} > E_{k}^{(S)}$.

Due to Corollary 1, a normal text tends to have a greater NRF score expected value than a stego-text.

Theorem 2. If the normal text $T_N$ and the stego-text $T_S$ have the same substitution list, then the more NRF unbalanced the substitution list is, the greater $E_{k}^{(N)}$ is than $E_{k}^{(S)}$.

Proof. Suppose that $0 < \delta \leq a < b < 1$, $a, b \in P_k$ (Note that $P_k = \{p_{k0}, p_{k1}, \ldots, p_{km_k-1}\}$ is the NRF set of the substitution set $k$ in the substitution list), if we can prove that when keeping any other value in $P_k$ unchanged and making a change to $a - \delta$ and $b$ change to $b + \delta$, $E_{k}^{(N)}$ increases more quickly than $E_{k}^{(S)}$, then it is apparent that Theorem 2 holds. In fact, due to the constant property of $q_{ki} (0 \leq i < m_k)$ within a certain substitution set and equation (6), these changes do not impact on the value of $E_{k}^{(S)}$ and then $E_{k}^{(S)}$, so we only have to examine if they make the value of $E_{k}^{(N)}$ increase. The increment of $E_{k}^{(N)}$ is

\[ \Delta E_{k}^{(N)} = \left[(a - \delta)(a - \delta) + (b + \delta)(b + \delta) - (a \cdot a + b \cdot b)\right] \quad \kappa(\delta^2 + (b + \delta)^2 - a^2 - b^2) = 2\kappa(\delta(a - b + \delta) > 0 \]

That is to say, the changes make $E_{k}^{(N)}$ become greater, which thereby makes $E_{k}^{(N)}$ be greater. As a result, Theorem 2 holds.

Due to Theorem 2, the more the substitution sets are of NRF unbalance, the greater the NRF score expected value of a normal text tends to be than that of a stego-text.

Now, it is time for us to consider the NRF score variances of a normal text $T_N$ and a stego-text $T_S$, namely $V_{k}^{(N)}$ and $V_{k}^{(S)}$, respectively. Also, assuming that they have the same substitution list $L = \{S_0, S_1, \ldots, S_{n-1}\}$, we define

\[ V_{k}^{(N)} = \frac{1}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{m_k-1} p_{ki} (c_{ij} - E_{k}^{(N)})^2 \]  
(9)

\[ V_{k}^{(S)} = \frac{1}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{m_k-1} q_{ki} (c_{ij} - E_{k}^{(S)})^2 \]  
(10)

Theorem 3. The NRF score variance of a stego-text $T_S$ is statistically greater than that of a normal text $T_N$ when they have the same substitution list of NRF unbalance and most of the substitution sets are of the same size, namely $V_{k}^{(S)} > V_{k}^{(N)}$.

Proof. In order to prove Theorem 3, we define an auxiliary variable $V_k^A$.

\[ V_{k}^{(A)} = \frac{1}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{m_k-1} p_{ki} (c_{ij} - E_{k}^{(S)})^2 \]  
(11)
We prove that $V^{(N)} < V^{(A)}$. Due to Corollary 1, $E^{(N)} > E^{(S)}$, we can assume $E^{(N)} = E^{(S)} + a$, $a > 0$, so

$$V^{(A)} = \frac{a}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} p_{ij} \left( c_{ij} - E^{(S)} + a \right)^2$$

$$= \frac{a}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} \left[ p_{ij} \left( c_{ij} - E^{(S)} \right)^2 + p_{ij} \cdot 2a \left( c_{ij} - E^{(S)} \right) + p_{ij}a^2 \right]$$

$$= V^{(N)} + 2a \cdot \left( \frac{1}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} p_{ij}c_{ij} - \frac{1}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} p_{ij}E^{(N)} \right)$$

$$+ \frac{4}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} p_{ij}a^2 = V^{(N)} + a^2$$

Thus we have $V^{(N)} < V^{(A)}$.

Then we prove that statistically $V^{(S)} \approx V^{(A)}$. Due to equations (10) and (11), we get

$$V^{(A)} - V^{(S)} = \frac{a}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} \left( p_{ij} - q_{ij} \right) \left( c_{ij} - E^{(S)} \right)^2 \approx \frac{a}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} \left( p_{ij} - q_{ij} \right) \left( q_{ij} \cdot p_{ij} - \kappa \cdot \frac{1}{m_j} \right)^2$$

$$= \frac{a}{n} \sum_{i=0}^{n-1} \sum_{j=0}^{m-1} \left( p_{ij} - q_{ij} \right)^3$$

(12)

In equation (12), as $E^{(S)}$ is algebraic average of $E^{(S)}(0 \leq i < n)$, when most of the substitution sets are of the same size $M$, in most cases, we have

$$E^{(S)} = \frac{1}{n} \sum_{i=0}^{n-1} E^{(S)} = \frac{1}{n} \sum_{i=0}^{n-1} \frac{1}{M} \kappa = \frac{\kappa}{M} = E^{(S)}$$

We substitute $E^{(S)}$ with $E^{(S)}$, keeping the equation statistically true.

Now, let us discuss the value of the sum $\sum_{j=0}^{m_j-1} \left( p_{ij} - q_{ij} \right)^3$. As $\sum_{j=0}^{m_j} p_{ij} = 1$, $q_{ij} = \frac{1}{m_j}$ is the algebraic average of $p_{ij}$, $j = 1, \ldots, m_j - 1$, the sum of the items $(p_{ij} - q_{ij})$ satisfies $\sum_{j=0}^{m_j-1} (p_{ij} - q_{ij}) = 0$. In the case of uniform distribution, the values $(p_{ij} - q_{ij})^3$ almost symmetrically distribute around 0. As a result, $V^{(A)} - V^{(S)} \equiv 0$ in the sense of statistics, that is $V^{(S)} \approx V^{(A)} = V^{(N)} + a^2$, namely $V^{(S)} > V^{(N)}$.

From Theorem 3, we can see that a stego-text tends to have a greater NRF score variance than a normal text.

**Corollary 2.** If the normal text $T_N$ and stego-text $T_S$ have the substitution list, the more the substitution list is of NRF unbalance, the greater $V^{(S)}$ is than $V^{(N)}$ statistically.

Due to Theorems 2 and 3, Corollary 2 holds obviously.

### 4. Detection scheme of substitution-based linguistic steganography

In Section 3, we closely examine the frequency characteristics of substitution elements in both the normal texts and stego-texts by relative frequency analysis and find some frequency properties that can be used to effectively distinguish between normal texts and stego-texts. In this section, we consider the application of these properties in the detection scheme.

When hiding information, the substitution-based linguistic steganography algorithm first scans the cover text to match the substitution elements in the substitution dictionary. If there is any part of text matching the element in the dictionary, then this part is replaced by one of the elements in the matched substitution set according to the hidden message. This procedure repeats until the hidden message is finished or the end of cover text is reached. The information hiding process is very straightforward. According to this process, we provide a detection scheme of substitution-based linguistic steganography as shown in Fig. 1.

In the detection scheme, there are two main steps: extracting classification features and classifying according to these features. In the classification step, an existent classifier such as SVM classifier (Chang and Lin, 2001) can be applied directly, so the main work is to design the first step. There are two procedures in the first step: one is processing the training corpus to get the NRF sets of Substitution Dictionary (SD) and the other is scanning the testing text $T$ to get the Substitution List (SL). Then the classification features are generated by using the NRF sets of SD and the SL. Here, the classification features are the NRF score expected value and variance value of the testing text. Due to Corollary 1 and the Theorem 3, a normal text tends to have a greater NRF score expected value and a less NRF score variance than a stego-text. Furthermore, the more the testing texts are of NRF unbalance, the greater the differences are between the statistical values in the case normal texts and those in the case of stego-texts. So in case that the testing texts are of considerably NRF unbalance, the detection scheme is promising.

![Fig. 1 – The detection scheme of substitution-based linguistic steganography. SD: substitution dictionary; SL: substitution list; $E_T$: NRF expected value of testing text $T$; $V_T$: NRF variance of testing text $T$.](image-url)
5. Detecting T-Lex system

T-Lex system (Winstein, 1999) is a famous synonym substitution-based linguistic steganography system proposed. In the system, as the substitution element is the synonym word, the substitution set and substitution dictionary become the synonym set and synonym dictionary. Apart from the name differences, the theorems described in Section 3 can be directly applied. In the section, as a demonstration we use our detection scheme to detect T-Lex system. In the following, the classification features that are used for detection are identified, and then the detection algorithm is given.

5.1. NRF score of synonym

We process a large training corpus to obtain the NRF sets of synonym dictionary, and then use the NRF sets and Equation (1) to calculate the corresponding NRF score sets. The NRF score sets are very critical in the detection. They are the basis of calculating the NRF scores of synonyms and then the expected values and variances of the scores.

5.2. NRF score expected value and variance

We denote the NRF expected value and variance of a testing text by \( \alpha \) and \( \gamma \) respectively. By the analogy of Equations (7)–(10) in the previous analysis, we can evaluate \( \alpha \) and \( \gamma \) values of a testing text. Due to the probability statistics sense, we use the NRF scores of the synonym words actually occurring in the testing text instead of the expected values calculated in these equations. Therefore, \( \alpha \) and \( \gamma \) can be evaluated using the occurring NRF scores \( c_i (0 \leq i < n) \) for the testing text by Equations (13) and (14).

\[
\alpha = \frac{1}{n} \sum_{i=0}^{n-1} c_i \quad (13)
\]

\[
\gamma = \frac{1}{n} \sum_{i=0}^{n-1} \left( \frac{c_i - \alpha}{\alpha} \right)^2 \quad (14)
\]

5.3. Detection algorithm

In the detection scheme shown in Fig. 1, we evaluate the NRF sets of synonym dictionary by training a large corpus, and then we can easily calculate the NRF score of each synonym word in synonym dictionary. The classification features are the NRF score expected value and variance of a testing text. These features can be evaluated by Equations (13) and (14) using the NRF scores of the substitution list for the testing text. The detection algorithm consists of the training algorithm and detecting algorithm, which are shown as Alg. 1 and Alg. 2.

Alg. 1 Description of training algorithm
1. For each text \( T \) in training corpus, do
   (1) Scan text \( T \) to match synonym words in synonym dictionary
   (2) Count the occurrences of each synonym word in text \( T \)

2. Calculate the occurrences of each synonym word in training corpus
3. Let \( \kappa = 1 \) and compute the NRF score of each synonym word in synonym dictionary using Equations (1) and (3).

Alg. 2 Description of detecting algorithm
1. For each word \( w \) in the testing text \( T \), do
   If word \( w \) is a synonym word, retrieve word \( w \)’s NRF score and push it into an array \( S \).
   2. Calculate the NRF score expected value and variance using the array \( S \) by Equations (13) and (14).
   3. Use an SVM classifier to classify the text \( T \) into normal text and stego-text, according to its NRF score expected value and variance.

Note that in our detection, we use an existent SVM classifier, but this is not necessary. One can use other classifier such as a Bayes classifier or even use some threshold values for low dimension classification feature vector such as in our case. However, we use an SVM classifier for the expandability of the classification feature vector.

5.4. Experiments and discussions

In our experiments, we use the raw text materials which are copied from the CD-ROM of English Classics 1000 published by Fudan University Press. We build a corpus of English literature containing thousands of text files as a basic corpus. We then build corpuses B-Corpus, C-Corpus, CD-Corpus, and S-Corpus from the basic corpus. The corpuses B-Corpus, C-Corpus, and S-Corpus consist of the literature works whose authors’ last names begin with “B”, “C” and “S” respectively, while CD-Corpus consists of the literature works written by Charles Dickens, who was the foremost English novelist of the Victorian era. The left literature works in the basic corpus compose another corpus named T-Corpus, which we use as the training corpus. Besides the training corpus, the SVM classifier requires two corpuses called SVM training set and SVM testing set, both of which have subsets named good set and bad set. The corpuses and their usages are list in Table 1.

Note that in Table 1, “Good Set” and “Bad Set” are sets of natural texts and stego-texts respectively, while “[X-Corpus]” means that the set of stego-texts consists of texts from X-Corpus processed by T-Lex system.

Figs. 2 and 3 show the classification feature distribution of the texts for the training and testing procedures of SVM. The \( x \)-axis and \( y \)-axis represent the NRF score expected value and variance of a text, which are denoted by \( \alpha \) and \( \gamma \). We can see
that the natural texts have greater $\alpha$ values and less $\gamma$ values than stego-texts in both training and testing procedures. As a result, the red start points representing stego-texts fall in the left-upper corner and the blue plus points representing natural texts fall in the right-lower corner. The classification feature distribution characteristics accords with the conclusions that we have drawn in the analysis in the Section 3.

When detecting, we only scan the first 5000 words in the texts. In fact, we just make use of the first several substitutions in the detection. The accuracies in case of using different substitution count (that is the length of the substitution list) in the detection are shown in Table 2. From the table, we can see that the detection accuracy rises on the whole as the substitution count increases and the accuracy soon gets considerable high even if the substitution count is still small. For example, when the substitution count is only 10, and 20, the detection accuracy exceeds 90%, 96% respectively.

In order to strictly assess the performance of our detection system, we apply the notion of precision and recall in addition to frequently used accuracy (Manning and Schütze, 2005). Table 3 shows the different parts of texts of our experiment when the substitution count is 90.

In Table 3, the true positive is $tp = 219$, the false positive is $fp = 5$, the false negative is $fn = 1$ and the true negative is $tn = 215$, then the precision, recall and accuracy are as follows:

$$\text{precision} = \frac{tp}{tp + fp} = \frac{219}{219 + 5} = 97.77\%$$

$$\text{recall} = \frac{tp}{tp + fn} = \frac{219}{219 + 1} = 99.55\%$$

$$\text{accuracy} = \frac{tp + tn}{tp + fp + fn + tn} = \frac{219 + 215}{219 + 1 + 5 + 215} = 98.64\%$$

We can see that the proposed detection system has high accuracy, precision and recall. This demonstrates that our detection scheme is of high performance. Additionally, according to Alg. 2, we can see that the detection time is nearly linear on the substitution count, so the detection scheme is also of high time efficiency.

So far, we have used the “pure” literature texts for our experiment. How about using the “mixed” or “balanced” texts? We carry out anther experiment applying the same NRF sets as before but using the balanced texts from English literature, academic reading and translated works, for both training and testing. The detection accuracies of different substitution count are show in the Fig. 4 in comparison to the previous results. We can see that detection accuracies obviously fall especially when the substitution count is small. This is as expected. In fact, how effectively our method works is mainly dependent on the how the NRF sets accord with the experimental texts. Since our current NRF sets are obtained from training a large number of literature texts, they accord better with the literature texts than the balanced texts. In spite of this, the detection accuracies is still promising, e.g. when the substitution count is not less than 50, the accuracies are above 90% and when the substitution is 100, the accuracy rich 95%. Additionally, from Fig. 4, it is shown that the difference between two lines gets smaller as the substitution count becomes greater.

<table>
<thead>
<tr>
<th>Substitution count</th>
<th>Success/total</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>411/440</td>
<td>93.41%</td>
</tr>
<tr>
<td>20</td>
<td>425/440</td>
<td>96.59%</td>
</tr>
<tr>
<td>30</td>
<td>427/440</td>
<td>97.05%</td>
</tr>
<tr>
<td>40</td>
<td>423/440</td>
<td>97.14%</td>
</tr>
<tr>
<td>50</td>
<td>422/440</td>
<td>95.91%</td>
</tr>
<tr>
<td>60</td>
<td>426/440</td>
<td>96.82%</td>
</tr>
<tr>
<td>70</td>
<td>427/440</td>
<td>97.05%</td>
</tr>
<tr>
<td>80</td>
<td>432/440</td>
<td>98.18%</td>
</tr>
<tr>
<td>90</td>
<td>434/440</td>
<td>98.64%</td>
</tr>
<tr>
<td>100</td>
<td>432/440</td>
<td>98.18%</td>
</tr>
</tbody>
</table>

Table 2 – The detection accuracies for pure literature texts.

<table>
<thead>
<tr>
<th>Substitution count</th>
<th>Success/total</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Detection</td>
<td>Stego-texts</td>
<td>219</td>
</tr>
<tr>
<td>Natural texts</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

Table 3 – Experiment results of our detection system for pure texts.
The pure texts and that for the balanced texts.

Plan of Suzhou (No. SYG201010).

of China (No. BK2010255), and the Scientific and Technical

the National Natural Science Foundation of China (No. 90818005),

This work was supported by the Major Research Plan of the
National Natural Science Foundation of China (No. 90818005),
the National Natural Science Foundation of China (No. 60903217),
the Natural Science Foundation of Jiangsu Province of
China (No. BK2010255), and the Scientific and Technical
Plan of Suzhou (No. SYG201010).

6. Conclusions and future work

As the development of linguistic steganography, the detection of linguistic steganography draws more and more attention of researchers all over the world. A few detection methods have been proposed and the corresponding detection tools have been developed. These in turn accelerate the research on more secure and advanced linguistic steganography methods.

In the paper, we first formally analyze the frequency characteristics of the normal texts and stego-texts, which we call RFA. Then the detection scheme of substitution-based linguistic steganography is proposed and the detection of synonym substitution-based linguistic steganography under this scheme is presented. Experimental results show that when using pure literature texts, the detection has a high accuracy even in the case that the substitution count is very small. The accuracies exceed 98% when the substitution count is not less than 80. While using the balanced texts, the accuracies fall as expected but the highest accuracy is still as high as 95%, showing that the method is also promising. Furthermore, the detection is also of high time efficiency. As a result, the proposed detection scheme is of practical significance.

In our current work, the substitution dictionary is assumed to be known and the secret information is assumed to be random binary string. Future work can focus on the detection in case that the substitution dictionary is unknown and the secret information is of certain kind of distribution.

Acknowledgements

This work was supported by the Major Research Plan of the National Natural Science Foundation of China (No. 90818005), the National Natural Science Foundation of China (No. 60903217), the Natural Science Foundation of Jiangsu Province of China (No. BK2010255), and the Scientific and Technical Plan of Suzhou (No. SYG201010).

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions.

REFERENCES


Topkara U, Topkara M, Atallah MJ. The hiding virtues of ambiguity: quantifiably resilient watermarking of natural

Fig. 4 — The comparison between detection accuracies for the pure texts and that for the balanced texts.