Covert Communication by Exploring Statistical and Linguistical Distortion in Text

Huanhuan Hu, Xin Zuo, Weiming Zhang(✉), and Nenghai Yu

University of Science and Technology of China, Hefei, China
zhangwm@ustc.edu.cn

Abstract. Most state-of-the-art text steganography algorithms are designed based on synonym substitution with the concern of simplicity and robustness. However, synonym substitution will cause some detectable impact on cover texts. In this paper, we propose an content-adaptive text steganography to minimize the impact caused by embedding process. We believe that synonym substitution will cause a hybird distortion consists of statistical distortion and linguistical distortion. We design a double-layered STC embedding algorithm (HSL) to minimize the distortion. Experiments results indicate that the security performance of HSL is better compared with traditional methods based on synonym substitution.

Keywords: Steganography · Synonym substitution
Statistical distortion · Linguistical distortion

1 Introduction

Encryption is a technique of protecting communications and its application can be found in all aspects of life. However, the garbled encrypted data will attract the attention from attackers which is not expected. The pursuit of steganography is behavioral safety. The existence of secret communication is hidden to avoid attackers attention. This does not mean that steganography is superior to encryption. The combination of these two methods can protect the secret information better [1].

There are many types of cover used in steganography such as text [2], image [3], audio [4] and video [5]. Because images, videos, etc. have redundant content and are not sensitive to modification, steganography developed for these cover has rich achievement. However, steganography that uses texts as cover develop slow as a result of the small content redundancy. However, the use of text data is increasing day by day with the rapid development of internet technology. This provides a natural environment for text steganography. The application prospect

An earlier version of this paper was presented at the 2nd IEEE International Conference on Data Science in Cyberspace.

X. Sun et al. (Eds.): ICCCS 2018, LNCS 11066, pp. 288–301, 2018.
https://doi.org/10.1007/978-3-030-00015-8_25

of text steganography is positive. Therefore, it is significant to design a secure steganography algorithm that is suitable for text cover.

Because of the high robustness and simplicity of the text steganography based on synonym substitution, this method is widely used [6–8]. However, the frequency distribution of synonyms in the cover text will change during the process of synonym substitution and causes statistical distortion. On the other hand, the meaning of a word is related to the content of the context. So the meanings of two synonyms in a particular context may not necessarily be the same and synonym substitution can cause linguistical distortion.

We will develop the rest of this paper as follows. In Sect. 2 we analyze the statistical distortion caused by synonym substitution. In Sect. 3 we introduce the method to estimate the linguistical distortion. The scheme for minimizing distortion of synonym substitution are elaborated in Sect. 4. Experimental results on resisting steganalyzers are shown for comparing with previous methods in Sect. 5. Section 6 gives the conclusion of this paper.

2 Statistical Distortion

2.1 Notation

To describe the proposed method more clearly, we give some notations as follows.

Definition 1. Embedding rate in steganography based on synonym substitution refers to the value of the number of bits embedded divided by the number of bits encoded by all synonyms in a text. The embedding capacity in synonym substitution steganography is decided by how many synonyms appear in a text file.

Definition 2. A synonym set is a word set which includes more than one words having similar meaning. The synonyms in the synonym set are order by the descending order of their frequencies which are derived from N-gram corpus. It's a open source corpus that can be downloaded from the Internet.

For example, (*Cow, Cattle*) is a synonym set that contains two synonyms and the frequency of Cow is bigger than Cattle.

To make the description in the rest of the paper simpler, we use a letter in lower case with a subscript and a superscript, take s_i^j for example, to represent a synonym. The corresponding synonym set is denoted as S_i, i.e., $s_i^j \in S_i$. $\|S_i\|$ denotes the number of synonyms in S_i. Herein, the subscript i is used to represent the position of the synonym appears in the text. The superscript j represents the order of the synonym in the corresponding synonym set. In this paper, if the logical expression $\boldsymbol{I}$ is true, we define the value of Iverson bracket $[\boldsymbol{I}]$ to be 1. Otherwise, the value of Iverson bracket $[\boldsymbol{I}]$ is defined to be 0.

Definition 3. A synonym sequence is defined as a sequence of synonyms. The synonyms are sorted in the increasing order of positions where they appear in the text. For example, if there are n synonyms in a text, the synonym sequence can be denoted as $(s_1^{j_1}, s_2^{j_2}, ..., s_n^{j_n})$.

2.2 Estimation of Statistical Distortion

Text steganography based on synonym substitution is widely used because it's simple and robust. In this kind of steganography, the sender and the receiver have the same synset. Different synonyms are encoded as different message bits. After we substitute some synonyms in cover text, the synonym sequence can represent the secret message. The semantics of the text remain almost unchanged. It's hard to distinguish if the text is modified. From this perspective, synonym substitution steganography is practical. However, experiments show that some statistical features of cover text will change during the process of synonym substitution.

As is known to all, synonyms appear at different frequencies in corpus. Rare synonyms refers to synonyms that hardly appear in corpus, while some synonyms appear frequently in corpus. We think that the message bits is random, **0** and **1** appear at the same probability in message. The number of **0** and the number of **1** in message are almost the same. As a result, rare synonyms appear at higher frequencies in stego texts than in cover texts. It can be utilized by attackers.

In this paper, we denote the synonym which has the highest frequency in a synonym set as MFS (Most Frequent Synonym). And the proportion of MFSs to all synonyms in a text is called Ratio of MFSs, denoted as R_M. We select 100 cover texts from wiki corpus. The size of these texts are 100 kB. We use Bsyn [9] to generate stego texts. The embedding rate is 0.5. We get R_M from these cover texts and stego texts and the results are shown at Fig. 1.

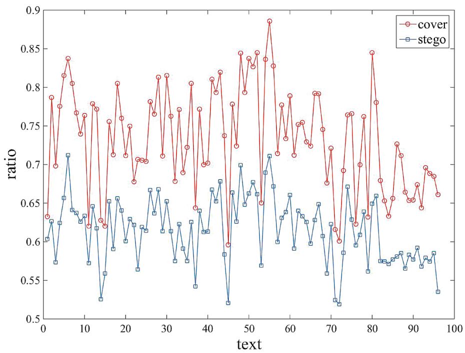

Fig. 1. Ratios of MFSs on the method Bsyn [9].

We can see from Fig. 1 that synonym frequencies will change if we embed message in the text by synonym substitution. It is regarded as statistical distortion which is not expected. As synonym substitution steganography is widely used, some researchers start research on text steganalysis. Text steganalysis can be used to detect the existence of secret messages in text files. Attackers may utilize text steganalysis tools to prevent covert communication. It's not we want if there is a need to transfer secret messages. However, most state-of-the-art synonym substitution steganography algorithms can not resist this kind of attack. In the perspective of statistical discrepancy, traditional synonym substitution steganography has room for improvement. The security performance of synonym substitution steganography need to be promoted and this is what we will do in this paper.

In this paper, the statistical distortion is estimated with the help of relative word frequency. If we substitute s_i^j with s_i^k, the statistical distortion is

$$SD(s_i^j, s_i^k) = \begin{cases} (log\frac{f(s_i^j)}{f(s_i^k)})^\alpha,\ if\ f(s_i^j) > f(s_i^k) \\ -(log\frac{f(s_i^k)}{f(s_i^j)})^\alpha,\ if\ f(s_i^j) <= f(s_i^k) \end{cases} \tag{1}$$

where the constant α is a parameter used to tune the sensitivity of the distortion to the frequency. $f(s_i^j)$ and $f(s_i^k)$ denote the frequencies of synonym s_i^j and s_i^k respectively, which are derived from the N-gram corpus.

3 Linguistical Distortion

3.1 Word to Vector

The research on how to represent words in the form of vectors has attracted much attention in recent years [10–12]. In [13], Bengio proposed a widely used model which could be used to estimate neural network language model. In Bengio's model, a feedforward neural network structure was adopted. With the help of a non-linear hidden layer and a liner projection layer, the performance of Bengio's model is somewhat satisfying. Many other later works learned how to build the neural network structure from Bengio's model. Another common structure of neural network language model was proposed in [14,15]. In this model, the word vectors were firstly learned using neural network with a single hidden layer. Then the word vectors were utilized to train the neural network language model. So the word vectors could be learned without constructing a complete neural network language model.

In [16], Tomas et al. proposed two new models for learning word vectors: Continuous Bag-of-Words (CBOW) and ssSkip-gram. CBOW model use the future and history words as input to predict the current word. Continuous Skip-gram model use the current word as input to predict words within a certain distance to the current word. In this paper, we use CBOW model to train word vectors. The architecture of CBOW is shown at Fig. 2.

It is noticed that the application of word vectors is not limited to simple linguistic regularities. Through applying simple addition and substraction operations on the word vectors, we can get the word vector of another word. For example, the result of word vector(King) - word vector(Man) + word vector(Woman) has a closest distance with the vector of the word "Queen" in vector space.

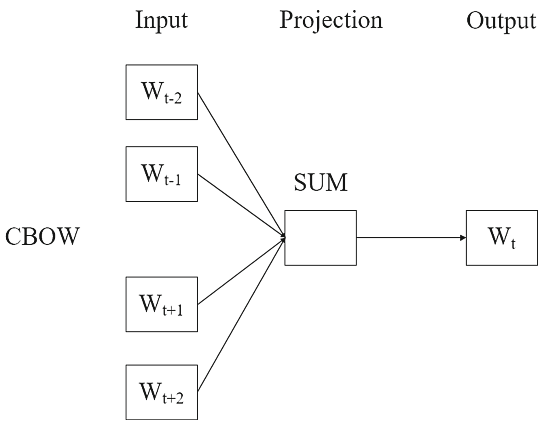

Fig. 2. The architecture of CBOW.

3.2 Estimation of Linguistical Distortion

A word can express different meanings in different contexts. Although the synonyms in a synset have the similar meaning, the substitution of synonyms will still cause a semantic mismatch in a specific context. It's necessary to quantify the linguistical distortion. In this paper, we utilize the word vectors to calculate the linguistical distortion caused by synonym substitution. Given a specific context, we can predict the current word with the help of word vectors. The prediction is realized by applying algebraic operation on the context word vectors and the result is in the form of a word vector. The weighted average of the context word vectors is the prediction result. And if the context is extracted from a text, we know the original word at this position in the text. We can calculate the distances between vector of the predicted word and vector of the original word. The distances is a measurement of how a word fits the specific context. In the remainder of this subsection, the calculation of linguistical distortion is given in detail.

For every synonym s_i^j, we can extract the N words before and after it as its context, denoted by $C_i = (c_{i,0}, c_{i,1}, ..., c_{i,2N-2}, c_{i,2N-1})$. The size N is called context window size. The vector representation of the context is $\boldsymbol{V}_i = (\boldsymbol{v}_{i,0}, \boldsymbol{v}_{i,1}, ..., \boldsymbol{v}_{i,2N-2}, \boldsymbol{v}_{i,2N-1})$. Since context words are often not of the same importance, we give different weights to the context words. $W_i = (w_{i,0}, w_{i,1}, ..., w_{i,2N-2}, w_{i,2N-1})$ is the weights of C_i. Given the context and the weights, we can predict the current word which may be not the same as the word appears in the text. The vector representation of the predicted word can be gotten by Eq. (2).

$$\boldsymbol{V}_{ip} = \sum_{k=0}^{2N-1} \boldsymbol{v}_{i,k} \times w_{i,k} \tag{2}$$

The vector representation of synonym s_i^j is denoted as $\boldsymbol{v}_i^j$. It is considered that the closer between $\boldsymbol{V}_{ip}$ and $\boldsymbol{v}_i^j$, the better s_i^j fits the context. In this paper, we choose cosine distance to measure the distance between two vectors. The reason of choosing cosine distance is given in Sect. 5. The cosine distance between $\boldsymbol{V}_{ip}$ and $\boldsymbol{v}_i^j$ is denoted as Cd_i^j.

$$cd_i^j = 1 - \frac{\boldsymbol{v}_i^j \cdot \boldsymbol{V}_{ip}}{|\boldsymbol{v}_i^j| \times |\boldsymbol{V}_{ip}|} \tag{3}$$

If we substitute s_i^j with s_i^k, the linguistical distortion is

$$LD(s_i^j, s_i^k) = \begin{cases} (log\frac{cd_i^k}{cd_i^j})^\beta,\ if\ cd_i^k > cd_i^j \\ -(log\frac{cd_i^j}{cd_i^k})^\beta,\ if\ cd_i^k <= cd_i^j \end{cases} \tag{4}$$

The parameter β is used to tune the proportion of linguistical distortion in total distortion.

4 The Proposed Scheme HSL

4.1 Preprocess the Synonym Sets

Considering the following two points, we firstly preprocess the synonym sets to guarantee that each synonym set contains 2 or 4 words to reduce the complexity of embedding algorithm.

(1) Little synonym set contains more than four synonyms;

(2) If $\|S_i\|$ =3 and synonyms in S_i are encoded into two bits, there will be wet elements in multi-layer STC [17].

The preprocess of synonym sets is conducted in the following ways: If a synonym set has three synonyms, remove the synonym with the lowest frequency. If a synonym set has more than four synonyms, remove the synonym with the lowest frequency until the synonym set contains only four synonyms.

We embed message by modifying the order j of the word s_i^j and substituting it with the corresponding synonym. To reduce embedding distortion with binary stego coding technology, we construct two binary cover sequences with the LSB (Least Significant Bit) and MSB (Most Significant Bit) of the order j of the synonym s_i^j. And if the synonym set includes only two words, the corresponding MSB will be empty.

For example, there is a synonym sequence $(s_0^0, s_1^0, s_2^1, s_3^2)$ and $\|S_0\| = 2, \|S_1\| = 2, \|S_2\| = 4, \|S_3\| = 4$. The two binary cover sequences are shown in Fig. 3.

4.2 Defining Distortion Function

In this paper, the distortion caused by substituting s_i^j with s_i^k is defined as

$$D(s_i^j, s_i^k) = SD(s_i^j, s_i^k) + LD(s_i^j, s_i^k) \tag{5}$$

Synonym	s_0^0	s_1^0	s_2^1	s_3^2
LSB	0	0	1	0
MSB	–	–	0	1

Fig. 3. Example of constructing cover sequences from synonym sequence.

We assume that the distortion of substitution of different synonyms is independent and the total embedding impact on cover text is the sum of distortion caused by every synonym substitution.

In the next subsection, we apply "minimizing distortion model for steganography" with Eq. (5) as the distortion metric. The distortion function $SD(s_i^j, s_i^k)$ means that replacing a word having higher frequency with one having lower frequency will introduce large costs, so such substitution will be limited, which can preserve the statistical character of texts. On the other hand, distortion function $LD(s_i^j, s_i^k)$ means substituting a synonym which fit the context better with another one will cause linguistical distortion. Therefore, by distortion metric (5), we take into account statistical distortion and linguistical distortion at the same time.

4.3 Applying Double-Layered STC

When the payload and distortion function are given, matrix embedding is a tool that can be employed to reduce the distortion during the process of embedding. Syndrome-trellis code (STC), a practical optimal code, can be utilized to hide message near the rate-distortion bound [18]. STC adopt convolutional code together with a Viterbi algorithm-based encoder for the purpose of minimizing the additive distortion. Previous works which use the framework of STC achieved satisfying performances [19–21]. Motivated by this, we developed a double-layered STC algorithm to implement the text steganography method proposed in this paper.

Suppose there is a synonym sequence $(s_1^{j_1}, s_2^{j_2}, ..., s_n^{j_n})$ in a cover text which is denoted as $\boldsymbol{x}$. $s_i^{j_i}$ belongs to synonym set $S_i = \{s_i^0, s_i^1, ...s_i^{n_i-1}\}$. It's possible that $S_i = S_k (i \neq k)$. We want to embed m bits of message into the text. After the message is embedded, the synonym sequence changes to $(s_1^{k_1}, s_2^{k_2}, ..., s_n^{k_n})$ which is denoted as $\boldsymbol{y}$. All possible value of $\boldsymbol{y}$ constitute a set which is denoted as $\boldsymbol{y}$. During the embedding process, the total distortion is formed as the sum of distortion caused by every synonym substitution.

$$D(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} D(s_i^{j_i}, s_i^{k_i}) \tag{6}$$

The probability of changing the synonym sequence from $\boldsymbol{x}$ to $\boldsymbol{y}$ is denoted as $\pi(\boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{x})$. The amount of bits can be sent is calculated by

$$H(\pi) = -\sum_{\boldsymbol{y} \in \boldsymbol{y}} \pi(\boldsymbol{y}) log \pi(\boldsymbol{y}). \tag{7}$$

Average distortion is calculated by

$$E_\pi(D) = \sum_{y \in \boldsymbol{y}} \pi(\boldsymbol{y}) D(\boldsymbol{x}, \boldsymbol{y}). \tag{8}$$

The task of embedding while trying to reduce the embedding impact is in the following form:

$$minimize\ E_\pi(D) \ \ subject\ to\ H(\pi) = m \tag{9}$$

According to the maximum entropy principle, the solution to Eq. (9) has a form of Gibbs distribution [22]

$$\pi(\boldsymbol{y}) = \frac{exp(-\lambda D(\boldsymbol{x}, \boldsymbol{y}))}{\sum_{\boldsymbol{z} \in \boldsymbol{y}} exp(-\lambda D(\boldsymbol{x}, \boldsymbol{z}))} \tag{10}$$

the parameter λ can be obtained by the constraint Eq. (9).

$$\pi(s_i^{k_i}) = p(s_i^{k_i} | s_i^{j_i}) = \frac{exp(-\lambda D(s_i^{j_i}, s_i^{k_i}))}{\sum_{z=0}^{n_i - 1} exp(-\lambda D(s_i^{j_i}, s_i^{z}))} \tag{11}$$

If $\|S_i\| = 2$, the possibility that the LSB of k_i is 0 is

$$p_{1i} = \pi(s_i^0) \tag{12}$$

If $\|S_i\| = 4$, the possibility that the LSB of k_i is 0 is

$$p_{1i} = \pi(s_i^0) + \pi(s_i^2) \tag{13}$$

According to [17], to reduce the embedding impact caused by embedding process, the payload of the first layer in double-layer STC is

$$m_1 = \sum_{i=1}^{n} -p_{1i} log p_{1i} - (1 - p_{1i}) log(1 - p_{1i}) \tag{14}$$

In the first layered embedding, the equivalent cover sequence is denoted as x_{1i} and

$$x_{1i} = [p_{1i} < 0.5], 1 \leq i \leq n \tag{15}$$

And the corresponding distortion metric of the element in the first layer cover sequence is

$$\rho_{1i} = |ln(\frac{p_{1i}}{1 - p_{1i}})| \tag{16}$$

The first layer stego sequence is denoted as y_{1i} $(1 \leq i \leq n)$, which is obtained by applying the STC to embed m_1 bits of message into the sequence x_{1i} $(1 \leq i \leq n)$.

The payload of the second layer binary sequence is

$$m_2 = m - m_1 \tag{17}$$

If $\|S_i\| = 4$, we can embed message into the MSB of the synonym. The possibility that the MSB of k_i is 0 is

$$p_{2i} = \begin{cases} \frac{\pi(s_i^0)}{\pi(s_i^0)+\pi(s_i^2)}, \, if \, y_{1i} = 0 \\ \frac{\pi(s_i^1)}{\pi(s_i^1)+\pi(s_i^3)}, \, if \, y_{1i} = 1 \end{cases} \tag{18}$$

where $i \in \{i| \, \|S_i\| = 4, i = 1, 2, ..., n\}$. In the second layered embedding, the equivalent binary cover sequence is denoted as x_{2i} and

$$x_{2i} = [p_{2i} < 0.5] \tag{19}$$

The corresponding distortion metric of the element in the second layer cover sequence is

$$\rho_{2i} = |ln(\frac{p_{2i}}{1 - p_{2i}})| \tag{20}$$

The second layer stego binary sequence is denoted as y_{2i}. When $i \in \{i| \, \|S_i\| = 4, si = 1, 2, ..., n\}$, y_{2i} is gotten by applying the STC to embed messages into x_{2i}. If $\|S_i\| = 2$, we set $y_{2i} = 0$. Finally, we set

$$k_i = 2y_{2i} + y_{1i} \tag{21}$$

Through comparing the difference between j_i and k_i, we know how to substitute the synonyms to get stego texts.

Note that the above double layered STC is different with that used for ± 1 embedding in images [17]. In image steganography, the embedding distortion is greatly influenced by the modification's amplitude and large modification amplitude means large distortion. Therefore ± 1 embedding overmatch two-layer LSB replacement for image steganography. When decomposing ± 1 embedding, some wet elements (disable elements) will appear in the second layer which may lead embedding failure, and thus we have to repeat the embedding process. However, in the proposed scheme, the cover element j is the order of a word in the synonym set. Large modification amplitude on j does not always mean large distortion. In fact, negative distortion may arise when a word with lower frequency is changed to one with high frequency. That's why we use two-layer LSB replacement instead of ± 1 embedding. For such cover on synonym substitution, we design a special double-layered STC to assign the payload to two layers of LSBs according to the distortion metric and the modification manner, which can achieve larger capacity than ± 1 embedding and will not yield wet elements. The details of the embedding and extraction procedures of the proposed method are described in Algorithm 1 and Algorithm 2 respectively.

5 Experiment Results

In this section, we first introduce the training process of the word to vector model. We adopt the CBOW model to train word vectors. The texts used as

Algorithm 1. Embedding Procedure

1: Get the synonym sequence $\boldsymbol{x}$ of cover text.
2: Calculate the distortion metric of $\boldsymbol{x}$ by using Eq. (5).
3: Determinate the value of λ by constraint Eq. (9).
4: Get the x_{1i} and the distortion metric of x_{1i}. Apply STC encoder to get y_{1i}.
5: Get the x_{2i} and the distortion metric of x_{2i}. Apply STC encoder to get y_{2i}.
6: Calculate k_i with Eq. (21) and generate $\boldsymbol{y}$.
7: Compare $\boldsymbol{x}$ and $\boldsymbol{y}$, replace the corresponding synonyms.

Algorithm 2. Extracting Procedure

1: Get the synonym sequence $\boldsymbol{y}$ of stego text.
2: Apply STC decoder to the LSB sequence of $\boldsymbol{y}$ get a part of message.
3: Apply STC decoder to the MSB sequence of $\boldsymbol{y}$ get the rest of message.

input are segmented from WIKI corpus. The size of input text ranges from 5 kB to 200 kB. The dimension of every word vector is 400-D. We set the context window size (parameter N) to 5 and abandon the words appeared less than 5 times during the training process. The synsets is extracted from Wordnet [23]. The synonyms in the synset are sorted in the descending order of their frequencies which are derived from N-gram corpus. In anti-detection experiments to evaluate the performance of different embedding methods, 5,000 texts are used as cover texts. The size of the text varies from 10 kB to 3000 kB. We try to guarantee these text files have a wide range of embedding capacities which is an effort to make the evaluation of the proposed method objective.

In Sects. 2 and 3, we give the calculation of statistical distortion and linguistical distortion in detail. The quantification of linguistical distortion is completed with the help of word vector. The distance between word vectors can be a measurement of linguistical distortion. Common vector distances include cosine distance and Euclidean distance. To determine the value of parameter α, β and find out which kind of vector distance can estimate linguistical distortion better, 1,000 texts was selected randomly from the WIKI corpus which are different from the 5,000 texts mentioned above. The stego texts are generated by HSL when α, β take different values and using different kind of vector distance. The embedding rate is 50%. The steganalysis tool is WFST. The detection result of steganalysis tool is displayed in Tables 1 and 2.

Table 1. Detection rate when α and β have different values (cosine distance)

Parameter	$\beta = 0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 1.5$
$\alpha = 0.0$	0.8319	0.8125	0.8286	0.8451
$\alpha = 0.5$	0.7958	0.7412	0.7882	0.8046
$\alpha = 1.0$	0.8033	**0.6769**	0.7426	0.7726
$\alpha = 1.5$	0.8154	0.7518	0.7274	0.7529

Table 2. Detection rate when α and β have different values (Euclidean distance)

Parameter	$\beta = 0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 1.5$
$\alpha = 0.0$	0.8319	0.8347	0.8551	0.8879
$\alpha = 0.5$	**0.7958**	0.8124	0.8485	0.8677
$\alpha = 1.0$	0.8033	0.8204	0.8467	0.8754
$\alpha = 1.5$	0.8154	0.8136	0.8652	0.8839

Compare Tables 1 and 2, we can know that the linguistical distortion can be better estimated by cosine distance. And HSL achieves optimal anti-detection when $\alpha = 1.0$ and $\beta = 0.5$. In the experiments later, we set $\alpha = 1.0$ and $\beta = 0.5$.

In Sect. 2, the ratios of MFSs are given to explain the statistical distortion caused by synonym substitution. In this section, the same experiment is implemented. The cover texts are identical with cover texts used in Sect. 2. The stego texts are generated by HSL (embedding rate = 0.5). Results are displayed in Fig. 4. Compare Figs. 1 and 4, we can see that the frequency distribution of synonyms in stego texts generated by HSL is closer to cover texts compared with stego texts generated by traditional method.

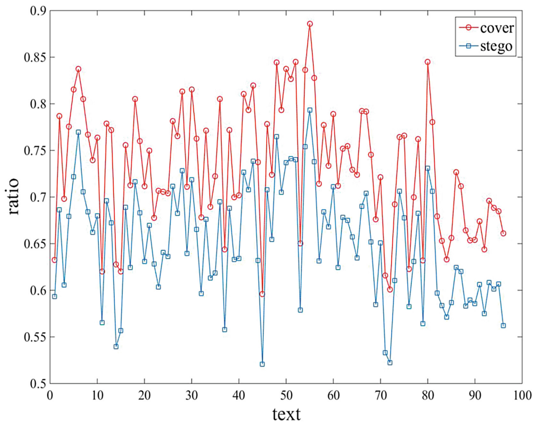

Fig. 4. Ratios of MFSs on the proposed method.

To further evaluate the security of HSL, two different steganalysis tools are used to detect stego text files produced by four methods: Bsyn, Tlex and Ctsyn [9] and HSL. Bsyn, Tlex and Ctsyn are all synonym substitution steganography. In Bsyn, the codewords of every synonym have the same length. Message are divided into many pieces of equal length. T-lex and Ctsyn don't limit the number of synonyms in synonym sets. T-lex uses WordNet to select synonyms with correct senses. Only the words appeared in the identical synonym set in WordNet database are grouped in a synonym set. Messages can be embedded into cover text as follow. First, encode the message letters with Huffman coding. Then, represent the encoded binary string in multi-base form. Finally, choose which synonym to appear in the text according to the multi-base form. Ctsyn constructs a binary tree for each synonym set with the synonyms as the leaves. Different synonyms represent message pieces of different lengths. For each steganographic method, two groups of stego texts are generated with two embedding rates 25% and 50%.

The first steganalysis tool is WFST. The detection results is displayed in Table 3. The second steganalysis tool is based on the context [24]. It's denoted as CST. The detection results is displayed in Table 4.

Table 3. Detection results of WFST

Embedding rate	Bsyn	Tlex	Ctsyn	HSL
25%	0.6907	0.6877	0.6802	**0.6131**
50%	0.8473	0.8359	0.8124	**0.7185**

Table 4. Detection results of CST

Embedding rate	Bsyn	Tlex	Ctsyn	HSL
25%	0.8995	0.8747	0.8864	**0.6639**
50%	0.9360	0.9214	0.9011	**0.7742**

From experiments results displayed above we can know that the security performance of HSL is better compared with other methods. HSL can resist steganalysis tools better by minimizing the designed distortion. It indicates the distortion quantification is reasonable.

6 Conclusions

In this paper, we analyze the distortion caused by synonym substitution from both the statistical and semantic perspectives. We apply minimal distortion model to synonym substitution steganography and design a double-layered embedding algorithm HSL to impact on cover texts during embedding process. Experiments show that HSL is more secure when attacked by different steganalysis tools compared with traditional synonym substitution steganography algorithms.

References

1. Shivani, Kumar, V., Batham, S.: A novel approach of bulk data hiding using text steganography. Procedia Comput. Sci. **57**, 1401–1410 (2015)
2. Huanhuan, H., Xin, Z., Weiming, Z., Nenghai, Y.: Adaptive text steganography by exploring statistical and linguistical distortion. In: 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC), pp. 145–150. IEEE (2017)
3. Denemark, T., Bas, P., Fridrich, J.: Natural steganography in JPEG compressed images. In: Electronic Imaging (2018)
4. Tayel, M., Gamal, A., Shawky, H.: A proposed implementation method of an audio steganography technique. In: 2016 18th International Conference on Advanced Communication Technology (ICACT), pp. 180–184. IEEE (2016)
5. Sadek, M.M., Khalifa, A.S., Mostafa, M.G.: Robust video steganography algorithm using adaptive skin-tone detection. Multimedia Tools Appl. **76**(2), 3065–3085 (2017)
6. Shirali-Shahreza, M.H., Shirali-Shahreza, M.: A new synonym text steganography. In: International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IIHMSP 2008, pp. 1524–1526. IEEE (2008)
7. Yuling, L., Xingming, S., Can, G., Hong, W.: An efficient linguistic steganography for chinese text. In: IEEE International Conference on Multimedia and Expo, pp. 2094–2097. IEEE (2007)
8. Muhammad, H.Z., Rahman, S.M.S.A.A., Shakil, A.: Synonym based malay linguistic text steganography. In: Innovative Technologies in Intelligent Systems and Industrial Applications, CITISIA 2009, pp. 423–427. IEEE (2009)
9. Xiang, L., Sun, X., Luo, G., Xia, B.: Linguistic steganalysis using the features derived from synonym frequency. Multimedia Tools Appl. **71**(3), 1893–1911 (2014)
10. Hinton, G.E.: Distributed representations (1984)
11. Hinton, G., Rumelhart, D., Williams, R.: Learning internal representations by back-propagating errors. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition 1 (1985)
12. Elman, J.L.: Finding structure in time. Cogn. Sci. **14**(2), 179–211 (1990)
13. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
14. Mikolov, T.: Language Modeling for Speech Recognition in Czech. Ph.D. thesis, Masters thesis, Brno University of Technology (2007)
15. Mikolov, T., Kopecky, J., Burget, L., Glembek, O., et al.: Neural network based language models for highly inflective languages. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2009, pp. 4725–4728. IEEE (2009)
16. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
17. Filler, T., Fridrich, J.: Minimizing additive distortion functions with non-binary embedding operation in steganography. In: 2010 IEEE International Workshop on Information Forensics and Security, pp. 1–6, December 2010
18. Filler, T., Judas, J., Fridrich, J.: Minimizing additive distortion in steganography using syndrome-trellis codes. IEEE Trans. Inf. Forensics Secur. **6**(3), 920–935 (2011)
19. Huang, F., Luo, W., Huang, J., Shi, Y.Q.: Distortion function designing for JPEG steganography with uncompressed side-image. In: Proceedings of the First ACM Workshop on Information Hiding and Multimedia Security. IH & MMSec 2013, pp. 69–76. ACM, New York (2013)

20. Li, B., Wang, M., Huang, J., Li, X.: A new cost function for spatial image steganography. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 4206–4210, October 2014
21. Zhao, Z., Guan, Q., Zhao, X.: Constructing near-optimal double-layered syndrome-trellis codes for spatial steganography. In: Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, pp. 139–148. ACM, New York (2016)
22. Fridrich, J., Filler, T.: Practical methods for minimizing embedding impact in steganography (2007)
23. Miller, G.A.: WordNet: a lexical database for English. Commun. ACM **38**(11), 39–41 (1995)
24. Chen, Z., Huang, L., Miao, H., Yang, W., Meng, P.: Steganalysis against substitution-based linguistic steganography based on context clusters. Comput. Electr. Eng. **37**(6), 1071–1081 (2011)