Copyright ⓒ 2021 KSII

An Improved Coverless Text Steganography Algorithm Based on Pretreatment and POS

Yuling Liu[1*], Jiao Wu[1], and Xianyi Chen[2]
[1] College of Computer Science and Electronic Engineering, Hunan University
Changsha, Hunan Province 410082 China
[e-mail: yuling_liu@126.com, lukywujiao@163.com]
[2] School of Computer and Software, Nanjing University of Information Science & Technology
Nanjing, Jiangsu Province 210044 China
[e-mail: 0204622@163.com]
[*]Corresponding author: Yuling Liu

Received November 15, 2016; revised August 4, 2018; revised September 24, 2018; accepted February 16, 2021; published April 30, 2021

Abstract

Steganography is a current hot research topic in the area of information security and privacy protection. However, most previous steganography methods are not effective against steganalysis and attacks because they are usually carried out by modifying covers. In this paper, we propose an improved coverless text steganography algorithm based on pretreatment and Part of Speech (POS), in which, Chinese character components are used as the locating marks, then the POS is used to hide the number of keywords, the retrieval of stego-texts is optimized by pretreatment finally. The experiment is verified that our algorithm performs well in terms of embedding capacity, the embedding success rate, and extracting accuracy, with appropriate lengths of locating marks and the large scale of the text database.

Keywords: Steganography, Coverless Text Steganography, Pretreatment, Chinese Character Component, Locating Mark, Part of Speech

This work was partially supported by the National Natural Science Foundation of China under Grant 61872134, Science and Technology Development Center of the Ministry of Education under Grant 2019J01020, Science and Technology Project of Transport Department of Hunan Province under Grant 201935 and Science and Technology Project of Changsha City. We express our thanks to Cuilin Wang who checked our manuscript.

http://doi.org/10.3837/tiis.2021.04.020 **ISSN : 1976-7277**

1. Introduction

Steganography is a hot spot in the field of multimedia content security. Different from encryption methods, steganography methods can imperceptibly embed secret message into multimedia carriers, such as image, video, audio, text, et al., therefore it is not easy to arouse suspicion [1].

Among the multimedia carriers, since text is the most frequently used in our daily life, we pay attention to text steganography, which includes three categories: text format based methods, text image based methods and linguistic steganography. The first method utilizes the text layout features, such as color, space and so on to hide information [2-3]. The biggest advantage of this method is large embedding capacity, whereas in terms of resistance to attacks, such as re-composition attacks and statistical analysis, this method performs poorly. Because texts can be regarded as binary images, it is feasible to embed information by utilizing the features of binary images. These methods are known as image-based text information hiding [4], which have very good imperceptibility, but cannot effectively resist re-composition attacks and Optical Character Recognition (OCR) attacks.

With development of natural language processing techniques, linguistic steganography methods have been proposed, which can be categorized as either generating methods or embedding methods. Generating methods employ text generation technologies to automatically generate texts with carrying secret message, such as steganography methods based on word lists [5] and grammar rules [6]. Because the generated cover texts are extremely unnatural, some schemes that generate texts of a special genre, such as notes [7], email addresses [8], poems [9], etc., have been proposed to obtain better performance. Embedding methods embed information by modifying the syntactic and semantic information in existing cover texts. There are lexical-level methods and sentence-level methods. In lexical-level methods, the secret message is embedded through the replacement of the vocabularies [10-12]. In sentence-level methods, the secret message is embedded by changing the structures of the sentences [13-15]. These methods have good robustness and imperceptibility. However, due to the limitations of natural language processing technologies, some methods are difficult to achieve. These methods cannot meet the various requirements of linguistic and grammar rules, and there are still some deviations in the language statistics [16].

The above steganography methods are not effective against steganalysis because they are usually achieved by modifying covers. To address this limitation, some researchers proposed a coverless steganography method for the first time [17-18]. This method emphasizes that any modification is not performed in the covers, and any additional cover is not needed. The development of big data technology enables the implementation of this method.

In terms of text, the main ideas of coverless steganography methods are driven by the secret message and using normal natural texts to transmit secret message. The key challenges include how to segment the secret information into specific keywords, how to generate the stego-texts by retrieving the text database, and how to locate the keywords while extracting the secret information. Ref. [18] presented a coverless text steganography algorithm based on Chinese Mathematical Expression. First, the algorithm directly generates a series of combinations of locating marks and keywords from the secret information. Some natural public texts are then retrieved as the stego-texts from the text database. By utilizing this method, a secret message can be transmitted to the receiver without any modification to any text. However, this method

has disadvantages, such as how to improve embedding capacity, the embedding success rate, and extracting accuracy as well as how to effectively retrieve the combinations of the locating marks and keywords in the text database. To solve this problem, an improved coverless text steganography algorithm based on pretreatment and POS is proposed.

The rest of this paper is organized as follows. Related work is introduced in Section 2. Section 3 introduces a framework of the scheme and algorithm details. Section 4 presents the experimental results. Finally, Section 5 is conclusions.

2. Related Work

In this section, we first present the background of coverless text steganography. Then, we introduce the Chinese Mathematical Expression and Part of Speech (POS) marking used in the proposed algorithm.

2.1 Coverless Text Steganography

Unlike conventional text steganography methods, coverless text steganography is to search the natural and public texts which include the secret information. There are many methods of retrieving information in plaintext and encrypted domains [19-20]. Besides, the sender does not transmit any assistant information to the receiver while using coverless text steganography.

In coverless text steganography, the secret message is first segmented into keywords, then insert a mark before each keyword, which is the locating mark. Subsequently, the text database is searched to obtain some relevant text documents, which are then transmitted to the receiver.

In the receiving side, the secret message can be extracted directly from the stego-texts via the locating marks and keywords segmentation methods, which figure out the problem of the previous text steganography algorithms based on Mimic needing to send a lot of extra information. Thus, since the receiver cannot extract the secret message using the conventional steganography method that extracts the modified features, the challenge of the coverless text steganography method is the retrieval of the stego-texts and the extraction of the secret message. The most accepted approach is the use of locating marks to label the locations of the keywords that compose the secret message. Ref. [21] proposed a linguistic coverless steganography method by utilizing the active learning based on named entity recognition, which employs entity recognition system to tag the location of the embedded information. Ref. [22] presented a coverless steganography algorithm based on the Chinese character encoding. In the method, the Chinese characters are transformed into a binary number to locate the secret message. Ref. [23] proposed a new algorithm of coverless steganography, which designed two mark selecting strategies. In [18], Chinese character components are selected as the locating marks, which will be explained in the following section.

2.2 Chinese Mathematical Expression

The Chinese Mathematical Expression was proposed for representing Chinese characters by Sun et al. in 2002 [24]. Chinese characters are divided into about 644 basic components. **Table 1** shows a selection of the basic components of Chinese characters.

Table 1. A part of the basic components of Chinese characters

1	2	3	4	5
一	乙	之	二	十
6	7	8	9	10
丁	厂	卜	人	八

The basic components are numbered, and there are six location relationships between them. The location relationships are left-right (lr), up-down (ud), left-down (ld), left-upper (lu), right-upper (ru), and whole enclosed (we).

For example, the Chinese character "从" consists of three parts, "人" and "一", in which they have two relationships: lr and ud. Therefore, from **Table 1** we can see, it will be calculated as "(9 lr 9) ud 1" by the Chinese mathematical expression.

2.3 POS Marking

POS marking is to label each word in a sentence with its appropriate POS [25], which is one of the key basic techniques of NLP. An example of a POS marking set is shown in **Table 2**. In the proposed algorithm, we construct a mapping set between POS and numbers by counting the POS after word segmentation. Using the POS to embed the number of keywords embedded in every stego-text can effectively improve the performance.

Table 2. An example POS tagging set

Numbers	Part of Speech	Subclass
1	n (noun)	nr (personal n.); ns (place n.)
2	a (adjevtive)	ad (adverbial adj.); an (nominal adj.)
3	v (verb)	vd (adverbial v.); vn (nominal v.)
4	q (quantifier)	Qv (quantifier verb); qt (quantifier time)
5	p (preposition)	pba ("Ba"); pbei ("Bei")

3. The Proposed Method

Mainly for Chinese characters, we propose an improved coverless text steganography method. In this section, we present the method, including full-text index building, locating marks selection, secret message segmenting into keywords, the combination of "locating mark" + "keyword" searching and secret message extraction. The proposed architecture is displayed in **Fig. 1**.

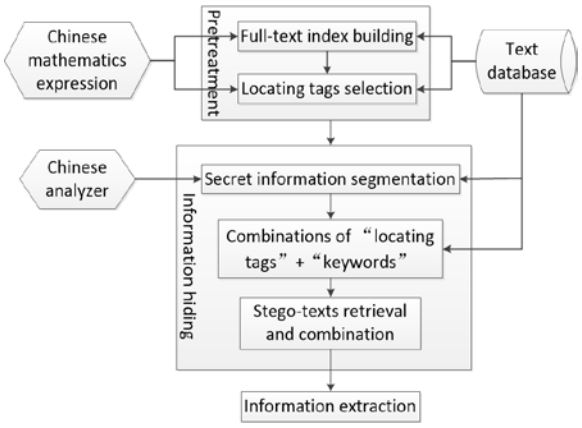

Fig. 1. Illustration of the proposed architecture

3.1 Notations

Some of the definitions and symbols are defined below.

- The number of Chinese character components is denoted as L.
- The text of database is denoted as T_i and its path is denoted as P_i.
- The keywords set is expressed as K, and the words set of T is denoted as W.
- The secret message is expressed as M .
- The combinations of the locating marks with keywords are denoted as $C = (L + W)$ or $C = (L + K)$.
- The stego-texts are denoted as S.

The POS of the words are expressed as O.

3.2 Pretreatment and Selection of Locating Marks

Pretreatment involves two parts: processing and counting the documents in the text database; and building the index to obtain the candidate set of locating marks and the mapping set between the digits and POS. In the proposed scheme, the Chinese character components are also utilized as the locating marks. The procedures are presented in the following.

(1) For any text in the database, acquire all the groups of the Chinese character components and words, denoted as $C = \{c_i, i = 1, 2,....., 644\}$, where $c_i = (l_i + w_i)$. Then calculate the POS of the words. The steps are in the following.

1.1) Partition the text and delete the stop words to acquire the list of words, denoted as $W=\{w_i, i =1, 2,.....\}$.

1.2) Except for the first word, for each w_i, take all the Chinese character components of the Chinese characters in w_{i-1} as the alternative locating marks of w_i, where $L_{wi} = (l_1, l_2,.....)$.

1.3) Check each l_i in L_{wi}. If l_i has not already been employed, it is used as the locating mark of w_i. If it has been employed, find w', where $l_i \in L_{w'}$. If the length of $L_{w'}$ is more than one, then set $L_{w'} = L_{w'} - l_i$ and take l_i as the locating mark of w_i; otherwise, set $L_{wi} = L_{wi} - l_i$. This step guarantees the uniqueness of each locating mark in every text and eliminates the needless locating marks in information extracting.

1.4) If there are the remaining locating marks, allocate these marks to the words without the marks in order. The above two steps can take advantage of the marks to improve embedding capacity.

1.5) Calculate and sort the POS of all the w_i.

(2) Construct two kinds of full-text indices of text database. The structures of the indices are like $l_i + k_i + p_i$ and $n_i + o_i + p_i$, where o_i is the POS and n_i is the list number of o_i, as shown in **Fig. 2**.

2.1) Calculate the items of the first index and construct the alternative locating mark table using the top N marks, denoted as $L = \{l_i, i = 1, 2,\ldots\ldots, N\}$.

2.2) Calculate the items of the second index and acquire the top POS set to construct the mapping relationship between the numbers and POS, denoted as $O_N=(o_1, o_2,\ldots\ldots)$, where the number i is mapped into o_i.

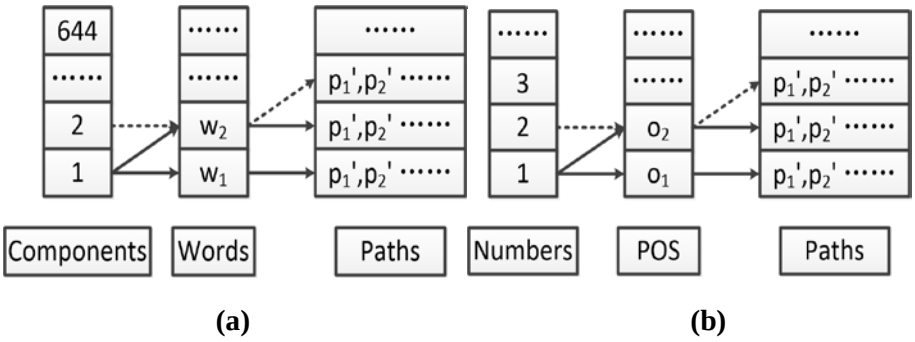

Fig. 2. Index structures of full text in the text database (a)"components"+"words"+"paths"; (b)"numbers"+"POS"+"paths."

3.3 Information Embedding

The partition of the secret message and the retrieval of the stego-texts are critical steps during the information embedding procedure, as shown in **Fig. 3**. It is important that the assembling of the keywords after partition and the corresponding locating marks can be successfully searched in the text database. Meanwhile, all combinations are selected in the same or similar texts as much as possible so that the sizes of the stego-texts are reduced.

The main steps involved are as below.

(1) Segment the secret message by calling word segmentation system NLPIR (http://ictclas.nlpir.org) and acquire the list of the keywords, denoted as $K = \{k_i, i = 1, 2,\ldots\ldots, M\}$.

(2) Generate the random sequence of the hidden marks from L by using the private key of the receiver, denoted as $L_1 = \{l_i, i = 1, 2,\ldots\ldots, M\}$.

(3) Embedding keywords: Search all assembling of L_1 and K. Acquire the stego-texts and the number list of the keywords embedded in the stego-texts. The stego-texts are denoted as $S_1 = \{t_i, i = 1, 2,\ldots\ldots, H_1\}$ and the number list is denoted as $I = \{i_i, i = 1, 2,\ldots\ldots, H_1\}$. If $H_1 = M$, then skip Step 4.

(4) Embedding numbers: According to the number list I, acquire the mapping POS set from O_N, denoted as $O_S = \{o_i, i = 1, 2, ..., H_1\}$. Generate the random number marks sequence by the private key of the receiver, denoted as $N = \{n_i, i = 1, 2,....., H_1\}$. After integrating N with O_S, if $o_i > 1$, search the integration by using the single keyword scheme and acquire $S_2 = \{t_i, i = 1, 2, ..., H_2\}$.

In the process of searching, if the integration $l_i + k_i$ has not been searched, then segment k_i to k_i' and k_i''. Add the two combinations, $l_i + k_i'$ and $l_{i+1} + k_i''$ and retrieve the new integration again. If k_i is an individual character, then the embedding of k_i fails.

(5) Acquire the final stego-texts S. S is made up of S_1 and S_2. Save S_1 and S_2 in different file names in order.

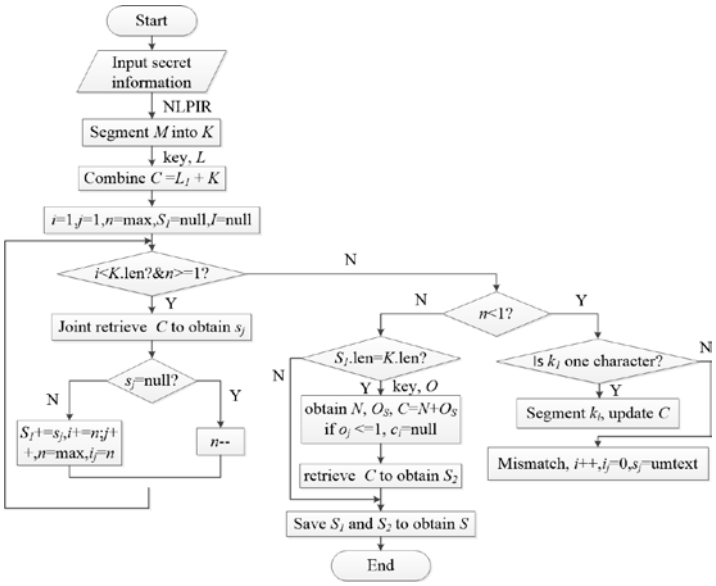

Fig. 3. A flowchart demonstrating the information hiding process

3.4 Information Extraction

Once the receivers accept the stego-texts, they can extract the embedded secret message by using the alternative locating marks set L, the mapping set O, and the receiver keys. A flowchart showing the information extracting steps is described in **Fig. 4**, and the main steps involved are described as below.

(1) Utilize the same pretreatment steps described in section 3.2 to acquire all assembling $l_i + w_i$ and order of the POS in every stego-text.

(2) Generate the random sequence of the locating marks $L_1 = \{l_i, i = 1, 2,....., H_1\}$ and calculate the random sequence $N = \{n_i, i = 1, 2,....., H_2\}$ by using the private key of the receiver.

(3) According to the file names, acquire S_1 and S_2 in order. If S_2 has content, extract the POS in S_2 according to the number marks of N and acquire the number sequences of the keywords embedded in S_1, denoted as $I = \{i_i, i = 1, 2,....., H_1\}$, where, if *ith* text is not in S_2, let $i_i = 1$. Then extract the corresponding keywords in S_1 according to I and L_1.

(4) Acquire the secret message by integrating all the keywords.

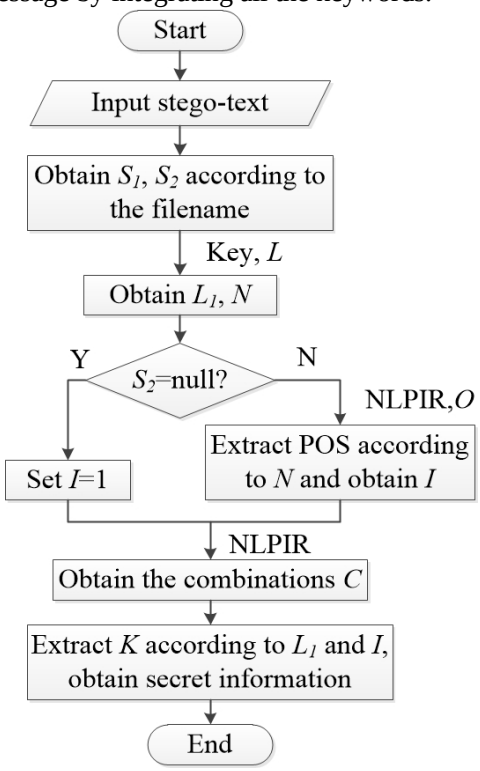

Fig. 4. A flowchart showing the information extraction process

4. Experiments and Analysis

We set up a Hadoop distributed cluster environment using the Linux operating system. The experimental data is collected mainly from articles, news, and Chinese text on the Internet. The text is divided into six categories: society, sport, tourism, education, culture, and military. There are around 10,000 articles in each category. The size of each article ranged from 1 to 6KB.

3,755 frequently-used Chinese characters that is part of the first class in GB2312 are picked as the reference for this paper. After deleting the stop words, the number of the frequently-used Chinese characters is dropped to 3,681. All the frequently-used Chinese characters are counted with the Chinese character components in the index items. The Chinese character components that could be combined with the most frequently-used Chinese characters are picked to construct the locating mark table. Finally, 200 locating marks are chosen.

Meanwhile, we make many experiments to calculate the numbers of keywords embedded in a text. According to the calculating results, one text can embed five keywords at best. Therefore, one mapping set between five frequently-used POS and the corresponding digits were constructed, namely $R = \{1 = n, 2 = a, 3 = v, 4 = q, 5 = p\}$, which is $O=\{n, a, v, q, p\}$. The results are described in **Fig. 5**, where the x-axis represents the length of secret message, the y-axis represents the number of texts, and the rectangles with different shapes are the numbers of keywords embedded in one text.

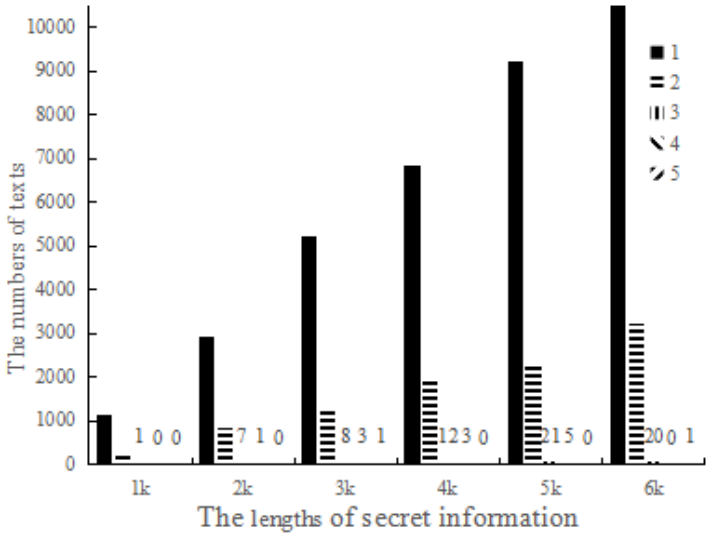

Fig. 5. Statistics on the number of keywords embedded in a text

4.1 An Example

So as to clearly demonstrate the information hiding process explained above, this section shows an example. With the secret message *M* being "无载体信息隐藏方法", the embedding process is as below.

First, segment *M* into the keywords sequence K = {无载体/信息隐藏/方法}. Then, generate a random sequence of locating marks $L_1 = \{9,5,7,4,.....\}$ by using the private key of the receiver from the locating marks table *L*. Obtain a random number mark sequence $N = \{3,5,4,1,2\}$. By combining the relevant elements in L_1 and *K*, search all the assembling C={9+无载体 5+信息隐藏 7+方法}. If the combination "9+无载体 5+信息隐藏 7+方法" cannot be searched in the text database, then continue to retrieve the combination "9+无载体 5+信息隐藏." If the combination "9+无载体 5+信息隐藏" also cannot be searched, then retrieve the combination "9+无载体". If the text t_1 is searched, set S1={t_1}, *I*={1}. Next, continue to retrieve the combination "5+信息隐藏 7+方法" in the text database. If text t_2 is searched, set the first stego-text set as S_1={t_1, t_2}, *I*={*1, 2*}. According to the mapping POS set O_N={*n, a, v, q, p*}, obtain the set O_s = {*null, a*}, where *null* represents that the text t_1 hides only one keyword, and the POS "*a*" represents that the text t_2 hides two keywords.

Second, we should hide the numbers of keywords to obtain the second stego-text set S_2. According to O_s = {*null, a*} and *N* = {*3,5,4,1,2*}, the combination *C* = {*5+a*} is searched in the text database. If text t_3 is retrieved, set S_2={t_3}.

Finally, combine sets S_1 and S_2 to obtain the final stego-text set *S* = {*0+*t_1, *1+*t_2, *a1+*t_3}.

4.2 Experiment and Results

So as to evaluate the performance of the proposed method, three performance measure

equations are calculated as below.

(1) The embedding success rate: It is an indicator to evaluate the performance of the embedding algorithm (denoted as σ):

$$\sigma = 1 - \frac{m_f}{m} \tag{1}$$

where m is the number of Chinese characters in the secret message and m_f is the number of individual characters that were not embedded successfully.

(2) Embedding capacity: This equation calculates the number of Chinese characters embedded in a text (denoted as ε , Chinese character/text):

$$\varepsilon = \frac{m}{m_s} \qquad \square \tag{2}$$

where m is the number of Chinese characters in the secret message and m_s is the number of the stego-texts.

(3) Extracting accuracy: So as to evaluate the performance of the extracting algorithm, the edit distance between the secret message and the extracted message is used to describe extracting accuracy (denoted as α):

$$\alpha = 1 - \frac{D}{L_m} \tag{3}$$

where D is the minimum number of editing operations transforming the extracted message into the secret message, which include the replacement of a character, the addition of a character, and the deletion of a character. L_m is the maximum length achievable between the extracted message and the secret message.

In the experiment, the secret message was randomly picked from the text database, of a 3KB size on average. It was independently embedded within the top 50, 100 and 150 marks in the locating marks table. The results are described in **Table 3**. Moreover, So as to evaluate the performance using different lengths of secret message, the secret message was randomly picked from the texts of 1KB, 2KB, 3KB, 4KB, 5KB, and 6KB sizes, with each group comprising 20 texts. The secret message was embedded within the top 200 marks of the locating marks table. The results are described in **Table 4**.

Table 3. Results of the experiment concerning different numbers of locating marks

Number of locating marks	Top 50	Top 100	Top 150
σ (Ref [23])	98.9%	100%	98.9%
σ (The proposed)	99.84%	99.86%	99.85%
α (The proposed)	89.79%	90.00%	89.96%
ε (Ref [18])	2.11	2.10	2.08
ε (The proposed)	2.24	2.24	2.25

Table 4. Results of the experiment concerning the different lengths of secret message

Length (KB)	1	2	3	4	5	6
σ (The proposed)	99.95%	99.94%	99.84%	99.90%	99.82%	99.94%
α (Ref [23])	71.25%	83.75%	91.88%	96.25%	98.13%	98.59%
α (The proposed)	94.20%	90.36%	90.47%	89.10%	88.33%	90.30%
ε (Ref [18])	2.09	2.15	2.08	2.10	2.11	2.11
ε (The proposed)	2.25	2.27	2.21	2.22	2.20	2.24

From the above results, it is clear that our scheme can achieve a high embedding success rate σ, as shown in (1). The reason for the failed embedding is that there are infrequent Chinese characters in the secret message. With the enlarge of the text database, the picking probability of the assembling can increase, and this improves the embedding success rate. Extracting accuracy α, as shown in (3), is very relevant to σ and the length of the secret message. As described in **Table 4**, with the secret message being the same length, the extracting accuracy α will be enhanced. Furthermore, we can find from **Table 3** and **Table 4** that embedding capacity ε, as shown in (2), has a slight change when different parameters are applied. The experiment results validate that the proposed algorithm has a little improvement in embedding capacity compared to the method used in [18].

4.3 Security Analysis

There are four kinds of attacks that place previous text steganography algorithms at risk: altering the format of texts (re-composition), recomposing the contents of texts (content attacks), text steganalysis algorithms based on text semantic (semantic steganalysis), and text steganalysis algorithms based on format statistical analysis (statistic steganalysis). A comparison of our method from other text steganography algorithms in terms of their resistance to attacks is shown in **Table 5**.

Steganography algorithms based on text format have been known for some time. For instance, in [3], the space characters were hidden in HTML/XML texts to embed information. These algorithms have a substantial embedding capacity, but they cannot resist re-composition attacks, content attacks, and statistic detection attacks. Since the proposed algorithm did not alter any text format due to the secret message, it is insensitive to re-composition attacks and can resist statistic detection attacks.

Image-based text steganography algorithms combine the features of binary images with text nature, such as dividing the text image into blocks and using the statistics of white and black pixels in every block [4]. Imperceptibility of these algorithms are good, but they cannot effectively resist re-composition attacks.

Linguistic steganography include generating algorithms and embedding algorithms, in which, the former generates new texts to carry the secret message and the later embeds information by altering the syntactic and semantic content of the existing texts. These linguistic steganography algorithms are all robust and have good imperceptibility. However,

due to the restriction of natural language processing technologies, they cannot meet some requirements which include rationality of parsing, accuracy of collocation and correctness of syntactic structure. Since the proposed method does not alter any text content due to the secret message, the stego-texts are natural texts. Thus the proposed algorithm can resist semantic steganalysis attacks and statistic steganalysis attacks.

Although our method can successfully resist the re-composition attack, semantic steganalysis attacks and statistic steganalysis attacks, it is not robust for content attacks. Because the method is based on the text content, once the natural text is deleted or tampered, the secret message extracted may be different from the original one. Thus, the method is sensitive to content attacks.

Table 5. Security comparison

Algorithms Attacks Methods	Re-composition	Content Attacks	Semantic Steganalysis	Statistic Steganalysis
Format-based [2-3]	Sensitive	Sensitive	Can resist	Cannot resist
Image-based [4]	Sensitive	Sensitive	Can resist	Can resist
Generation Method [5-6]	Insensitive	Sensitive	Can resist	Cannot resist
Embedding Method [10-15]	Insensitive	Sensitive	Cannot resist	Cannot resist
Proposed Method	Insensitive	Sensitive	Can resist	Can resist

The proposed method improves on that which was proposed in [18].

(1) Embedding capacity: The proposed method can hide multiple keywords. Moreover, during the information embedding process, the stego-texts that only embed single keyword do not need to embed the number of keywords. Thus, it is useful for improving embedding capacity.

(2) Extracting accuracy: In the pretreatment phase, the uniqueness of the locating mark in every text is guaranteed. Thereby, the ambiguity of the locating marks in the information extracting process can be eliminated. Thus, extracting accuracy is improved.

(3) Embedding success rate: By selecting all the Chinese character components of every Chinese character in every word as the locating marks, each text uses 644 locating marks. Meanwhile, the number of the keywords in stego-texts is mapped to POS. Thus, the embedding success rate is improved.

5. Conclusion

This paper have proposed an improved coverless text steganography algorithm based on pretreatment and POS. The Chinese character components are used as the locating marks. POS are used for embedding the number of keywords to increase embedding capacity. Meanwhile, the segmentation of the secret message and the retrieval process of the stego-text are optimized by pretreatment. Using natural texts instead of altering the existing texts to embed information,

the proposed algorithm can effectively resist all kinds of steganalysis attacks, and it can improve the security of transmitting the secret message. However, this study does have its limitations, which can be mitigated as follows:

(1) Increasing embedding capacity: The embedding capacity of the method can be improved by expanding the text database, using recommended similar keywords in the generation of the secret message, and optimizing the retrieval method of the combinations.

(2) Improving security: In future studies, we will consider converting the secret message into the similar keywords to improve the security of the algorithm.

References

[1] I. J. Cox and M. L. Miller, "The first 50 years of electronic watermarking," *Journal of Applied Signal Processing*, vol. 2002, no. 2, pp. 126-132, 2002. Article (CrossRef Link)

[2] S. Zhang, Z. Yao, X. Meng, and C. C. Liu, "New digital text watermarking algorithm based on new-defined characters," in *Proc. of IEEE International Conference Symposium on Computer, Consumer and Control (IS3C)*, pp. 713-716, 2014. Article (CrossRef Link)

[3] I. S. Lee and W. H. Tsai, "Secret communication through web pages using special space codes in HTML files," *International Journal of Applied Science & Engineering*, vol. 6, no. 2, pp. 141-149, 2008. Article (CrossRef Link)

[4] M. Wu and B. Liu, "Data hiding in binary image for authentication and annotation," *IEEE Transactions on Multimedia*, vol. 6, no. 4, pp. 528-538, 2004. Article (CrossRef Link)

[5] K. Maher, "Texto". [Online]. Available: ftp://ftp.funet.fi/pub/crypt/steganography/texto.tar.gz

[6] M. Chapman and G. I. Davida, "Hiding the hidden: a software system for concealing ciphertext as innocuous text," in *Proc. of the 1st International Conference on Information and Communication Security*, pp. 335-345, 1998. Article (CrossRef Link)

[7] A. Desoky, "Notestega: Notes-based steganography methodology," *Information Security Journal: A Global Perspective*, vol. 18, no. 4, pp. 178-193, 2009. Article (CrossRef Link)

[8] E. Satir and H. Isik, "A huffman compression based text steganography method," *Multimedia Tools and Applications*, vol. 70, no. 3, pp. 2085-2110, 2014. Article (CrossRef Link)

[9] Y. B. Luo, Y. F. Huang, F. F. Li and C. C. Chang, "Text steganography based on Ci-poetry generation using markov chain model," *KSII Transactions on Internet and Information Systems*, vol. 10, no. 9, pp. 4568-4584, 2016. Article (CrossRef Link)

[10] C. Y. Chang and S. Clark, "Practical linguistic steganography using contextual synonym substitution and vertex color coding," *Computational Linguistices*, vol. 40, no. 2, pp. 403-448, 2010. Article (CrossRef Link)

[11] B. Feng, Z. H. Wang, D. Wang, C. Y. Chang, and M. C. Li, "A novel, reversible, Chinese text information hiding scheme based on lookalike traditional and simplified Chinese characters," *KSII Transactions on Internet and Information Systems*, vol. 8, no. 1, pp. 269-281, 2014. Article (CrossRef Link)

[12] L. Y. Xiang, Y. Li, W. Hao, P. Yang, and X. B. Shen, "Reversible natural language watermarking using synonym substitution and arithmetic coding," *Computers, Materials & Continua*, vol. 55, no. 3, pp. 541-559, 2018. Article (CrossRef Link)

[13] B. Murphy and C. Vogel, "The syntax of concealment: reliable methods for plain text information hiding," *Security, Steganography, and Watermarking of Multimedia Contents IX*, vol. 6505, pp. 1-12, 2007. Article (CrossRef Link)

[14] B. Murphy and C. Vogel, "Statistically-constrained Shallow Text Marking: Techniques, Evaluation Paradigm and Results," *Security, Steganography, and Watermarking of Multimedia Contents IX*, vol. 65050Z, 2007. Article (CrossRef Link)

[15] C. Y. Chang and S. Clark, "Linguistic steganography using automatically generated paraphrases," in *Proc. of Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 591-599, 2010. Article (CrossRef Link)
[16] P. Meng, L. Hang, W. Yang, and Z. Chen, "Attacks on Translation Based Steganography," in *Proc. of IEEE Youth Conference on Information, Computing and Telecommunication (YC-ICT)*, pp. 227-230, 2009. Article (CrossRef Link)
[17] Z. Zhou, H. Sun, R. Harit, X. Chen, and X. Sun, "Coverless image steganography without embedding," in *Proc. of International Conference on Cloud Computing and Security (ICCCS)*, pp. 123-132, 2015. Article (CrossRef Link)
[18] X. Chen, H. Sun, Y. Tobe, Z. Zhou, and X. Sun, "Coverless information hiding method based on the Chinese mathematical expression," in *Proc. of International Conference on Cloud Computing and Security (ICCCS)*, pp. 133-143, 2015. Article (CrossRef Link)
[19] Y. L. Liu, H. Peng, and J. Wang, "Verifiable eiversity ranking search over encrypted outsourced data," *Computers, Materials & Continua*, vol. 55, no. 1, pp. 37-57, 2018. Article (CrossRef Link)
[20] X. Shen, F. Shen, Q. S. Sun, Y. Yang, Y. H. Yuan, and H. T. Shen, "Semi-Paired Discrete Harshing: Learning Latent Hash Codes for Semi-Paired Cross-view Retrieval," *IEEE Transactions on Cybernetics*, vol. 47, no. 12, pp. 4275-4288, 2017. Article (CrossRef Link)
[21] H. Sun, R. Grishman, and Y. Wang, "Active learning based named recognition and its application in natural language coverless information hiding," *Journal of Internet Technology*, vol. 18, no. 2, pp. 443-451, 2017. Article (CrossRef Link)
[22] X. Chen, S. Chen, and Y. Wu, "Coverless information hiding method based on the Chinese character encoding," *Journal of Internet Technology*, vol. 18, no. 2, pp. 313-320, 2017. Article (CrossRef Link)
[23] Y. Wu and X. Sun, "Text coverless information hiding method based on hybrid tags," *Journal of Internet Technology*, vol. 19, no. 3, pp. 649-655, 2018. Article (CrossRef Link)
[24] X. Sun, H. Chen, L. Yang and Y. Y. Tang, "Mathematical representation of a Chinese character and its applications," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 16, no. 6, pp. 735-747, 2002. Article (CrossRef Link)
[25] A. Ekbal and S. Saha, "Simulated annealing based classifier ensemble techniques: application to part of speech tagging," *Information Fusion*, vol. 14, no. 3, pp. 288-300, 2013. Article (CrossRef Link)

Yuling Liu is currently an Associated professor in the College of Computer Science and Electronic Engineering at Hunan University, China. She was Visiting scholar at UMASS Lowell in 2016. She received the Ph. D degree in Computer Science from Hunan University, China, in 2008. Her research interests include network and information security, information hiding based on big data, text analysis.

Jiao Wu received her M. S. degree in Computer Science from Hunan University, China, in 2008.

Xianyi Chen received his Ph.D. degree at Hunan University in 2014. He was a visiting scholar at University of The University of North Carolina at Pembroke in 2018. He is currently an Associate professor in Nanjing University of Information Science and Technology and an Executive editor of Journal on Big Data. His research interests include Information security of AI, Information hiding based on Big data, digital forensics, watermarking in encrypted domain, etc.