Blind Linguistic Steganalysis against Translation Based Steganography

Zhili Chen*, Liusheng Huang, Peng Meng, Wei Yang and Haibo Miao

1. NHPCC, School of CS. & Tech., USTC, Hefei 230027, China
2. Suzhou Institute for Advanced Study, USTC, Suzhou, 215123, China
zlchen3@ustc.edu.cn

Abstract. Translation based steganography (TBS) is a kind of relatively new and secure linguistic steganography proposed. It takes advantage of the "noise" created by automatic translation of natural language text to encode the secret information. Up to date, there is little research on the steganalysis against this kind of linguistic steganography. In this paper, a blind steganalytic method, which is named natural frequency zoned word distribution analysis (NFZ-WDA), is presented. This method has improved on a previously proposed linguistic steganalysis method based on word distribution which is targeted for the detection of linguistic steganography like nicetext and texto. The new method aims to detect the application of TBS and uses none of the related information about TBS, its only used resource is a word frequency dictionary obtained from a large corpus, or a so called natural frequency dictionary, so it is totally blind. To verify the effectiveness of NFZ-WDA, two experiments with two-class and multi-class SVM classifiers respectively are carried out. The experimental results show that the steganalytic method is pretty promising.

1 Introduction

Enlightened by the word distribution analysis (WDA) linguistic steganalysis method proposed by Chen *et al.* [1], this paper presents an improved method which can blindly distinguish natural texts, machine translated texts and stego texts that generated by translation based steganography (TBS) [2][3][4]. The key idea is to examine the distribution characteristics of words that are in the same natural frequency zone (NFZ). When using a machine translator to translate texts from one language to anther, as the translator uses words somehow in a mechanical way, the translated texts have an inherent structural style determined by the machine translator. Therefore, the stego texts generated by TBS have a mixed structure style that determined by all the translators used. Similarly, the people's writing texts, which we call natural texts, also have an inherent structural style.

The paper evaluates the potential of distinguishing the structural styles of different classes of texts. Our basic observation is that the structural style can be well represented by the distributions of words in the same NFZs. In order to

characterize the inherent structural style differences of different classes of texts, we mainly focus on the investigation of distribution characteristics of different NFZs of words. In our work, we first attribute the words in the text being analyzed into different NFZs according to their natural frequencies, which are the word frequencies obtained from a large corpus. Next, we find the positions of the words in the same NFZ and calculate the average and the variance of the distances between neighboring words. Finally, we use the distance averages and variances of all the NFZs to form the classification feature vector representing the structural style, based on which we use a SVM classifier to distinguish between different classes of texts.

The NFZ-WDA method has improved on the previous WDA method by introducing the notion of NFZ and refining the word distribution characteristics. The WDA method analyzes the testing texts entirely based on the texts themselves, while the NFZ-WDA method applies in the analysis a frequency criterion, which provides a more correct direction. Additionally, the refinement of the word distribution characteristics preserves more structural information. As a result, the improved method makes it more possible to effectively analyze the stego texts generated by TBS.

The organization of the paper is as follows: Section 2 briefly covers the basic operations of the TBS algorithm and the previous steganalytic methods against TBS. Section 3 focuses on the description of the blind linguistic steganalysis method, NFZ-WDA. In Section 4, we present the results of our steganalytic experiments and give some related analysis. In Section 5, there are some discussions about NFZ-WDA. Finally, conclusions are presented in Section 6.

2 Related Work

2.1 Translation based Steganography

Compared to traditional linguistic steganography methods such as nicetext and texto, translation based steganography (TBS), which was introduced by Grothoff *et al.*, is a novel and relatively secure method. TBS hides information in the "noise" created by automatic translation of natural language text. The key idea of TBS is "When translating a non-trivial text between two natural languages, there are typically many possible translations. Selecting one of these translations can be used to encode information." [2] Because there are frequent errors in legitimate automatic translated texts, it is difficult for a computer to distinguish the additional errors inserted by an information hiding algorithm and the normal noise associated with translation. Therefore, steganalysis against TBS is a challenging work.

There are two versions of TBS, lost in translation (Lit)[2] and lost in just the translation (LiJtT)[3]. Lit works as follows. At first, the sender picks up a cover text in the source language, which does not have to be kept secret and can be obtained from public sources. Then, the sender translates the source text to target language sentence by sentence using several translators and encodes

the hidden messages in this process by selecting one proper translator for each source sentence.

The work flow of LiJtT is illustrated in Fig. 1. LiJtT has improved on Lit to one that allows the receiver recovering the hidden message using only stego texts and a secret key. LiJtT works as follows. First, the sender generates multiple translations for a given cover text and uses a secret key which is shared between the sender and the receiver to hash each translated sentence into a bit string. Second, the lowest h bits of the hash string, referred to as header bits, are interpreted as an integer $b \geq 0$ and then the sentence whose lowest $[h+1\ ,h+b\ ]$ bits match with the bit-sequence to be encoded is selected. Finally, when the receiver receives a stego text, he breaks the received text into sentences, applies a keyed hash to each sentence and interprets the lowest $[h+1,\ h+b]$ bits of each hash string as the next b bits of the hidden message.

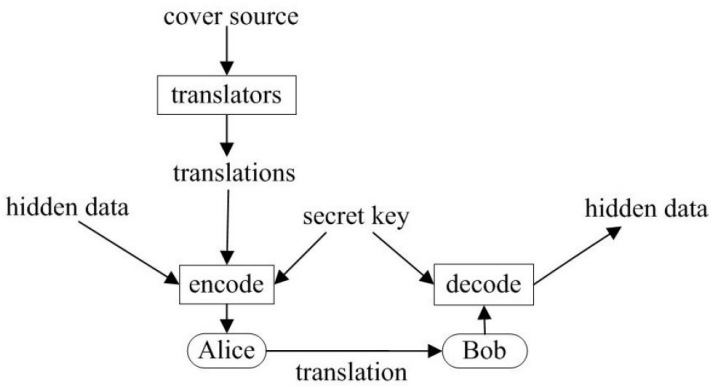

Fig. 1. Work Flow of LiJtT [3]

2.2 Previous Linguistic Steganalysis Methods of TBS

The steganalytic method on TBS presented by Meng *et al.* [5] needs to know the machine translator (MT) set and the cover text language. As the translator set is the private key of TBS, the method has to seek the possible candidate translator set before the steganalysis. This weakens the commonality of the method. Furthermore, the steganalytic process of the method has to translate the cover text two times by every translator, which may be too expensive for large-scale deployment.

The steganalytic method proposed by Meng *et al.* [6] no longer needs to know the MT set used by the TBS encoder in the steganalysis. The method is based on the fact that there are fewer high-frequency words in stego texts than

in normal texts. By defining the feature vector related to the frequencies of the high-frequency words, the authors use a two-class SVM classifier to detect the stego texts and normal texts.

The latter steganalytic method described above has achieved a promising steganalysis, but it still suffers some drawbacks. Firstly, it is not a completely blind steganalysis, that is, it still needs some prepared resources, namely the sets of high-frequency words and n-grams, which seem to be more or less related to the TBS. Next, the countermeasure of this steganalytic method still seems plausible, especially by using only one-to-one words and 2-grams in the TBS. Finally, the detection accuracy needs to be improved, particularly when the text size is less than 20KB. The steganalytic method proposed in this paper can properly overcome these drawbacks.

3 Blind Linguistic Steganalysis Method

3.1 Previous WDA Method

In the WDA method [1], the spread degree (SD) of a word is defined as the variance of its positions in the testing text. Then the average and variance of the SDs of words in the testing text are used to form the classification feature vector. Finally, a two-class SVM classifier is applied to classifying the testing text to normal texts and stego texts.

Though the WDA method summarizes the structural information of the testing text so much that a lot of detailed information are lost, it still works quite effectively for detection of linguistic steganography methods such as nicetext [7], texto [8] and Markov chain based [9]. Preserving the correct syntax and coherent semantics well, the TBS generates more natural-like texts. As a result, the stego texts of TBS are more difficult to analyze. The WDA method has no effect on the steganalysis of TBS as we will see in Section 4.

3.2 The Improvement of WDA Method

The steganalytic method proposed in this paper, NFZ-WDA, has improved on the WDA method in three aspects as follows.

First, the natural frequency dictionary is used in NFZ-WDA while no language resources are used in WDA. The natural frequency dictionary is used as a guide for the NFZ partition and provides a frequency criterion by which the distribution features of words with a certain range of natural frequencies can be calculated.

Second, positions of the words in the same NFZ are used as a whole to calculate the word distribution features in NFZ-WDA while only the positions of the same word are used in WDA. By doing this, NFZ-WDA can abstract more invariant features about the word distribution.

Third, the distance average and variance of each NFZ are used to form the classification feature vector in NFZ-WDA while in WDA, the average and variance of the spread degrees of words in the testing text are used. NFZ partition makes the description of word distribution is more accurate and detailed.

3.3 NFZ-WDA Method

Definitions Before introducing the NFZ-WDA method, some definitions have to be clarified as follows.

Natural Frequency is a word's general frequency in the natural texts. The natural frequency of a word represents the occurring probability of the word in the natural texts.

Natural Frequency Dictionary is the set of natural frequencies of a certain word dictionary. It can be evaluated by processing a large corpus and calculating the word frequencies.

Natural Frequency Zone (NFZ) is the set of words of a certain natural frequency range. That is, the words in a NFZ have approximative natural frequencies.

In this paper, we attribute the words to different NFZs according to their natural frequencies and investigate the distribution characteristics of words in each NFZ.

Text Formulation As done in the paper [1], we formalize a text as follows.

$$T = \{w_0, w_1, w_2, ..., w_{n-1}\} \tag{1}$$

Where w_i, $0 \leq i \leq n-1$ is the $(i+1)$ th word of the text.Then, the word position of word w_i is defined as.

$$l_i = \frac{i}{n} \tag{2}$$

Obviously, $0 \leq l_i \leq 1$.

In the NFZ-WDA method, we make use of the natural frequency dictionary and assign words to different NFZs according to their natural frequencies. Given the natural frequency set of the text T, which is obtained by retrieving the natural frequency dictionary,

$$F = \{f_0, f_1, f_2, ..., f_{n-1}\} \tag{3}$$

and the maximal natural frequency f_{max} in the natural frequency dictionary, the NFZs with equal size are formulated as

$$Z_k = \{w_i | kL \leq f_i < (k+1)L\}, k = 0, 1, ..., K-1 \tag{4}$$

Here L is the NFZ size and $K = \lceil \frac{f_{max}}{L} \rceil$ is the count of NFZs.

After having formulated the NFZ, the text T can be regarded as the constitution of words from all the NFZs. Suppose that the text T contains the words from NFZ Z_k with n_k times, we have

$$\sum_{k=0}^{K-1} n_k = n \tag{5}$$

The word position set of NFZ Z_k can be denoted by

$$L(Z_k) = \{l_0^{(k)}, l_1^{(k)}, l_2^{(k)}, ..., l_{n_k-1}^{(k)}\} \tag{6}$$

subject to $l_0^{(k)} < l_1^{(k)} < l_2^{(k)} < ... < l_{n_k-1}^{(k)}$. Particularly, let $l_{-1}^{(k)} = 0$ and $l_{n_k}^{(k)} = 1$.

Let Z denote the set of NFZs, L denote the set of $L(Z_k)$. That is to say

$$Z = \{Z_k | k = 0, 1, ..., K-1\} \tag{7}$$

$$L = \{L(Z_k) | k = 0, 1, ..., K-1\} \tag{8}$$

We can represent the text T in another form, where

$$T = \langle Z, L \rangle \tag{9}$$

Classification Feature Vector Each class of texts has a unique structural style. The more accurately we can describe the structural style, the more effectively we can distinguish from different classes of texts. Our key observation is that the structural style of a certain class of texts can be well represented by its word distribution characteristics. So, if we can accurately describe the word distribution characteristics of a text, we can identify its class with a high efficiency.

In order to measure the distribution of words, we first define the distance of words w_i and w_j in the text T as

$$d_{ij} = d_{ji} = |l_i - l_j| \tag{10}$$

Then, we define in the NFZ Z_k the average and variance of the distances of neighboring words, which are denoted by α_k and γ_k, as

$$\alpha_k = \frac{1}{n_k + 1} \sum_{i=0}^{n_k} d_{i,i-1}^{(k)} = \frac{1}{n_k + 1}(1 - 0) = \frac{1}{n_k + 1} \tag{11}$$

$$\gamma_k = \frac{1}{n_k + 1} \sum_{i=0}^{n_k} (d_{i,i-1}^{(k)} - \alpha_k)^2 \tag{12}$$

Here, $d_{i,i-1}^{(k)} = |l_i^{(k)} - l_{i-1}^{(k)}|$ denotes the distance of $w_i^{(k)}$ and $w_{i-1}^{(k)}$, which are the $(i+1)$ th and i th words in the NFZ Z_k. The borders are defined as $d_{0,-1}^{(k)} = |l_0^{(k)} - l_{-1}^{(k)}| = |l_0^{(k)} - 0| = l_0^{(k)}$ and $d_{n_k,n_k-1}^{(k)} = |l_{n_k}^{(k)} - l_{n_k-1}^{(k)}| = |1 - l_{n_k-1}^{(k)}| = 1 - l_{n_k-1}^{(k)}$.

Finally, we define the classification feature vector Γ as

$$\Gamma = \{(\alpha_k, \gamma_k) | k = 0, 1, ..., K-1\} \tag{13}$$

Therefore, vector Γ contains the word distribution information of words in each NFZ. As the size of NFZ decreases, it can describe the text structural style to an inch.

Method Description In the steganalysis, there are normally three main processes employed: training, testing and classifying. We apply the following procedure to both training and testing processes to abstract the classification feature vector from the processed text.

Step 1: Word Position Computation. The analyzer reads the given text T, parsing it, splitting it into words and obtains the word set T in the form of Eq. (1). Then the analyzer computes the word position of each word in the given text using Eq. (2). See Alg. 1.

Step 2: NFZ Partition. The analyzer loads the natural frequency dictionary, retrieves the natural frequencies of the words in the text T and attributes them to different NFZs according to their natural frequencies. In this step we can get Z_k and then the corresponding word position set $L(Z_k)$, among which $k = 0, 1, ..., K$. See Alg. 2.

Step 3: Distance Average and Variance Computation. Using each word position set $L(Z_k)$, the analyzer computes the distances between any two neighboring words in each NFZ Z_k, namely gets the $d_{i,i-1}^{(k)}$, where $i = 0, 1, ..., n_k$. Then, the analyzer computes the distance average and variance of each NFZ Z_k according to Eq. (11) and (12). See Alg. 3.

Algorithm 1 Word Position Computation

Splits the text T into words and gets the total word count n
for all $w_i \in T$ **do**
  $l_i \leftarrow \frac{i}{n}$
end for

Algorithm 2 NFZ Partition

Loads natural frequency dictionary and retrieves the natural frequencies of the words in the text T to get natural frequency set F.
$Z_k \leftarrow \emptyset$
$L(Z_k) \leftarrow \emptyset$
for all $w_i \in T$ **do**
  $k \leftarrow \lfloor \frac{f_i}{L} \rfloor + 1$
  $Z_k \leftarrow Z_k \cup \{w_i\}$
  $L(Z_k) \leftarrow L(Z_k) \cup \{l_i\}$
end for

Going through all these three steps described above, the analyzer converts the given text T to a classification feature vector Γ as formulated in Eq. (13). The vector Γ is then used as the exclusive basis for the text classification.

Having introduced the classification feature extraction algorithms, we move to the description of the whole steganalytic system. Fig. 2 shows the NFZ-WDA

Algorithm 3 Distance Average and Variance Computation

for all $Z_k \in Z$ **do**
  $\alpha_k \leftarrow \frac{1}{n_k+1}$
  $\gamma_k \leftarrow 0$
  for $i = 0$ to n_k **do**
    $d^{(k)}_{i,i-1} \leftarrow |l_i - l_{i-1}|$
    $\gamma_k \leftarrow \gamma_k + (d^{(k)}_{i,i-1} - \alpha_k)^2$
  end for
  $\gamma_k \leftarrow \sqrt{\frac{\gamma_k}{n_k+1}}$
end for

system framework. As described previously, the framework mainly includes training, testing and classifying three parts. The former two parts are constituted of the same three-step classification feature extraction, while the latter part is an existent SVM classifier [10]. The arrowhead represents the flow of data, among which the dashed line arrowhead indicates that the training process can be omitted if the classification model has already been prepared. The thick dashed rectangle indicates the whole steganalytic system.

While analyzing, both the training and testing texts go through the three-step classification feature extraction, resulting in training and testing feature set. Then the training feature set is used for the training of the SVM classifier and generating of the classification model. The testing feature set is used for the classification. The classification results indicate the steganalytic conclusions.

In the system framework, apart from the training and testing texts, there is only a natural frequency dictionary used, which is applied to NFZ partition in the process of classification feature extraction. Furthermore, the natural frequency dictionary is obtained from a large corpus and it has nothing to do with TBS. As a result, the NFZ-WDA System is a totally blind steganalytic system.

4 Experiments and Analysis

In our experiments, texts were translated from German to English using the LiT Prototype, with no semantic substitution, no article and preposition replacement enabled and no "badness threshhold" [2]. The translator set of LiT includes Systran, Google, Promt translators.

We have built the training and testing text sets from natural language texts, machine translated texts and stego-texts. The natural language texts which are in English were extracted from a corpus of 1000 classical English novels, the machine translated texts were generated using Systran, Google and Promt translators and the German language texts which are used as cover texts and the source language texts of translation came from the Europarl corpus [11]. All the experimental texts are in the form of text segment with a size of about 20KB. Our detector utilizes only the first thousands of bytes indicated by a text size parameter, e.g., if the text size is 5KB, it means that the detector only uses the first 5KB text of each experimental text of size 20KB.

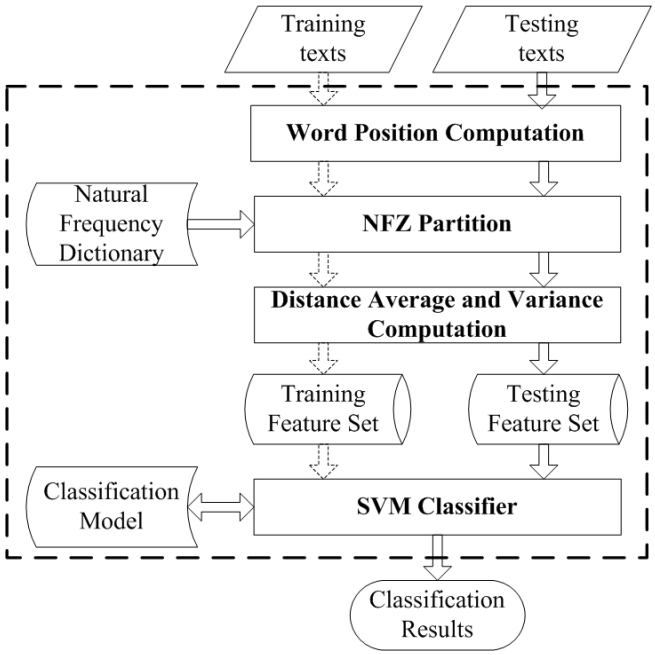

Fig. 2. NFZ-WDA System Framework

In order to measure the detection, we use some rates that are defined as follows.

$$\text{False Rate} = \frac{\text{number of non-stego texts identified as stego}}{\text{total number of non-stego texts}}$$

$$\text{Missing Rate} = \frac{\text{number of stego texts identified as non-stego}}{\text{total number of stego texts}}$$

$$\text{Accuracy Rate} = \frac{\text{number of testing texts identified as their true type}}{\text{total number of testing texts}}$$

In the first experiment, we use WDA method [1] to distinguish between stego-texts and natural language texts or one kind of the machine translated texts. Texts of about 20KB size are used. Tab. 1 shows the experimental results. Obviously, The WDA method has little effect on the steganalysis of TBS.

Table 1. Experimental Results of WDA. The "Train" and "Test" columns show the numbers of texts used in the training and testing for both classes. "FR", "MR" and "AR" are short for "False Rate", "Missing Rate" and "Accuracy Rate" respectively.

Text Size	Class-1	Class-2	Train	Test	FR(%)	MR(%)	AR(%)
	Natural	Stego	60/60	172/202	72.67	31.19	49.73
20KB	Promt	Stego	60/60	211/202	58.77	39.11	50.85
(abt 3300	Google	Stego	60/60	185/202	75.14	13.37	57.11
words)	Systran	Stego	60/60	210/202	84.29	12.38	50.97

Then, in order to verify the feasibility and effectiveness of the improved steganalytic method, we have designed yet two experiments. The first one is the experiment distinguishing between stego-texts and natural language texts or one kind of the machine translated texts using a two-class SVM classifier. The second is the experiment distinguishing among the stego-texts, natural language texts and all kinds of machine translated texts using a multi-class SVM classifier.

For both experiments, we use a natural frequency dictionary which is generated using the written English texts from the British National Corpus (BNC) [12]. In the dictionary, the maximal natural frequency f_{max} is found to be 6187927. We then let the NFZ size $L = 10$ and get the count of the NFZs is $K = \lceil \frac{f_{max}}{L} \rceil = \lceil \frac{6187927}{10} \rceil = 618793$. We use the texts with sizes varying from 5KB to 20KB, or with word counts varying from about 800 to about 3300. As an NFZ implies two classification features, namely distance average and variance of words in the NFZ, the dimensionality of the theoretical classification feature vector is very large. But as the word count of each testing text is limited, the dimensionality of the actual classification feature vector is not more than twice of the word count, that is hundreds or thousands of classification features is

actually used to describe the structural style of each class of texts. Fig. 3 and Fig. 4 show the distributions of the first 10 NFZs' distance average features and distance variance features for each text class. From these pictures, we can image that each text class has a inherent, unique structural style represented by the distribution of words via which we can detect different classes of texts.

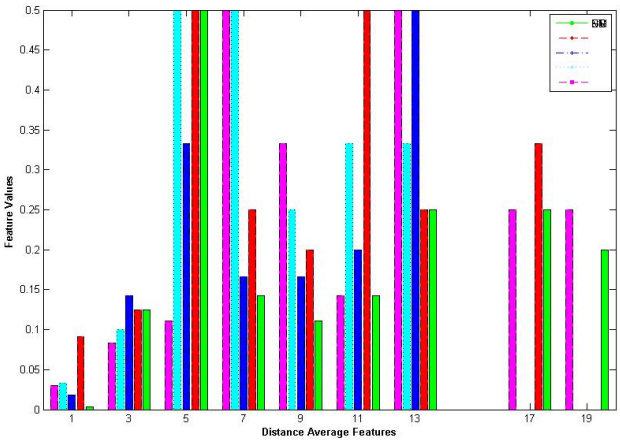

Fig. 3. The distributions of the first 10 distance average features for each text class. The X axis value $2k-1$ represents the distance average feature of kth NFZ which is evaluated by α_k, where $k = 1, 2, ..., 10$.

Tab. 2 and Tab. 3 show the experimental results of distinguishing both between two classes of texts and among five classes of texts. On the whole, both tables show that the proposed analytic method is highly promising.

In Tab. 2, the distinguishing between the stego texts and the natural texts is of very high accuracy, almost 100%, no matter how the text size is. This means that we can easily differentiate stego texts from natural language using our method. The accuracy of distinguishing between stego texts and one kind of the machine translated texts is around 93% when the text size is 5KB and above 97% when the text size is not less than 10KB, which is pretty ideal.

In Tab. 3, when the text size is 5kB, the total detection accuracy is 91.22% and the detections of natural texts and the machine translated texts as non-stego texts have accuracy rates of above 90%, but the detection of stego texts has a poor accuracy of 75.74%. This may be caused by the improper generation of the classification model in the training process of the SVM classifier. The detection accuracies of both non-stego texts and stego texts increase as the text size increases when the text size is 10kB or above. The total detection accuracies

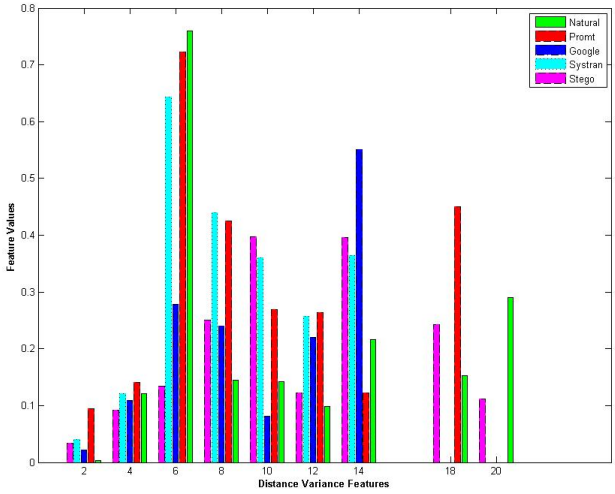

Fig. 4. The distributions of the first 10 distance variance features for each text class. The X axis value $2k$ represents the distance variance feature of kth NFZ which is evaluated by γ_k, where $k = 1, 2, ..., 10$.

Table 2. Experimental Results of NFZ-WDA using Two-class SVM classifier. The "Train" and "Test" columns show the numbers of texts used in the training and testing for both classes. "FR", "MR" and "AR" are short for "False Rate", "Missing Rate" and "Accuracy Rate" respectively.

Text Size	Class-1	Class-2	Train	Test	FR(%)	MR(%)	AR(%)
5K (abt 800 words)	Natural	Stego	60/60	172/202	0.00	0.00	100.00
	Promt	Stego	60/60	211/202	9.48	3.96	93.22
	Google	Stego	60/60	185/202	0.00	10.89	94.32
	Systran	Stego	60/60	210/202	2.38	10.40	93.69
10K (abt 1600 words)	Natural	Stego	60/60	172/202	0.58	0.00	99.73
	Promt	Stego	60/60	211/202	1.90	3.47	97.34
	Google	Stego	60/60	185/202	0.00	0.50	99.74
	Systran	Stego	60/60	210/202	0.95	0.00	99.51
15K (abt 2500 words)	Natural	Stego	60/60	172/202	0.00	0.00	100.00
	Promt	Stego	60/60	211/202	3.32	0.50	98.06
	Google	Stego	60/60	185/202	0.00	0.00	100.00
	Systran	Stego	60/60	210/202	0.00	0.00	100.00
20K (abt 3300 words)	Natural	Stego	60/60	172/202	0.00	0.00	100.00
	Promt	Stego	60/60	211/202	1.90	0.00	99.03
	Google	Stego	60/60	185/202	0.00	0.00	100.00
	Systran	Stego	60/60	210/202	0.00	0.00	100.00

are 97.65%, 98.88% and 99.69% respectively when the text size is 10KB, 15kB and 20kB, which is also ideal.

Table 3. Experimental Results of NFZ-WDA using Multi-class SVM classifier. The "Train" and "Test" columns show the numbers of texts used in the training and testing for each class.

Text Size	Class	Train	Test	Non-stego(%)	Stego(%)	Accuracy(%)
5K (abt 800 words)	Natural	60	172	100.00	0.00	91.22
	Promt	60	211	91.00	9.00	
	Google	60	185	96.76	3.24	
	Systran	60	210	94.29	5.71	
	Stego	60	202	24.26	75.74	
10K (abt 1600 words)	Natural	60	172	100.00	0.00	97.65
	Promt	60	211	98.10	1.90	
	Google	60	185	97.30	2.70	
	Systran	60	210	97.62	2.38	
	Stego	60	202	4.46	95.54	
15K (abt 2500 words)	Natural	60	172	100.00	0.00	98.88
	Promt	60	211	98.58	1.42	
	Google	60	185	99.46	0.54	
	Systran	60	210	98.57	1.43	
	Stego	60	202	1.98	98.02	
20K (abt 3300 words)	Natural	60	172	100.00	0.00	99.69
	Promt	60	211	99.53	0.47	
	Google	60	185	100.00	0.00	
	Systran	60	210	100.00	0.00	
	Stego	60	202	0.99	99.01	

5 Discussions

Before we complete the presentation of NFZ-WDA steganalytic method against TBS, there are some discussions about this method as follows.

First, as we have pointed out, the method uses none of the information related to TBS. Apart from the training and testing texts, the only required resource for this method is the natural frequency dictionary, which can be obtained from any one of the large enough corpuses. So it is a totally blind steganalysis against TBS.

Second, the main underlying basis of this method is that each kind of texts has a unique structural style and we use the distributions of words in all the NFZs to describe this style. The countermeasure of this method will be a goal hard to reach, for the modification of the word distributions in all NFZs is extremely difficult.

Third, the experimental results show that the steganalytic method has achieved high detection accuracies, which is superior to the previous steganalytic methods against TBS.

Finally, NFZ-WDA is very simple to achieve and can be easily applied in other natural language, e.g., the application of NFZ-WDA to the authorship attribution of Chinese novels proves to be successful in our initial experiments.

In fact, as the texts generated by certain linguistic steganography method usually have a unique structural style, the proposed method can also be used to analyze texts generated by other linguistic steganography methods, such as nicetext, texto, spammimic, mimicry and so on. Besides, as the texts written by different men also have a unique structural style, NFZ-WDA method also can be used to distinguish texts written by different people. So it can also be applied in the steganalysis of TBS method that uses manual translations and other research areas like authorship analysis and text forensics. The verification of these applications is our future work.

6 Conclusions

In this paper, an improved method for the steganalytic method proposed by Chen *et al.* [1] is presented. The new method called natural frequency zoned word distribution analysis (NFZ-WDA) is used as a blind steganalytic method against translation based steganography (TBS). Our contributions are summarized as follows.

1) We have found a weakness in TBS: the texts generated by different translators have their inherent, unique structural style that is determined by the translator itself.

2) We have found a highly effective way to describe the unique structural style of certain class of texts, which not only can be used in the steganalysis against TBS, but also can be used in the steganalysis of other linguistic steganography methods and other research areas such as authorship analysis and text forensics.

3) We have proposed to use a two-class SVM classifier to distinguish between stego-texts and natural language texts or one kind of the machine translated texts, and to use a multi-class SVM classifier to distinguish among the stego-texts, natural language texts and all kinds of machine translated texts. Both detection accuracies are pretty high and increase as the text size increases.

Stego texts generated by TBS are basically preserved syntactically correct and semantically coherent. The difficulty of detecting TBS depends on many factors such as how many translators TBS used, which translators, the source language and the target language. The previous analytic methods need to use some information more or less related to the steganographic method itself and the detection accuracies still need to be improved. This paper presents a totally blind steganalytic method against TBS. The experimental results show that our steganalytic method is highly promising.

7 Acknowledgement

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 (Nos. 60903217 and 60773032), the China Postdoctoral Science Foundation funded project (No. 20090450701) and the Scientific and Technical Plan of Suzhou (No. SYG201010).

References

1. Chen Zhili, Huang Liusheng, Yu Zhenshan, Li Lingjun, Yang Wei. A Statistical Algorithm for Linguistic Steganography Detection Based on Distribution of Words. In Proc. of ARES 2008: 558-563.
2. Grothoff Christian, Grothoff Krista, Alkhutova Ludmila, Stutsman Ryan and Atallah Mikhail. Translation-Based Steganography. In Proc. of Information Hiding 2005: 221-233.
3. Stutsman Ryan, Grothoff Christian, Atallah Mikhail and Grothoff Krista. Lost in Just the translation. In the Proc. of ACM symposium on Applied computing 2005: 338-345.
4. Grothoff Christian, Grothoff Krista, Stutsman Ryan, Alkhutova Ludmila, and Atallah Mikhail. Translation-based steganography. Journal of Computer Security, 17(3): 269-303, 2009.
5. Meng Peng, Huang Liusheng, Yang Wei, and Chen Zhili. Attacks on Translation based teganography. In Proc. of IEEE Youth Conference on Information, Computing and Telecommunication 2009: 227-230.
6. Meng Peng, Huang Liusheng, Chen Zhili, Yang Wei and Hu Yuchong. STBS: A Statistical Algorithm for Steganalysis of Translation-Based Steganography. Accepted by Information Hiding 2010.
7. Chapman Mark. Hiding the Hidden: A Software System for Concealing Ciphertext as Innocuous Text. http://www.NICETEXT.com/NICETEXT/doc/ thesis.pdf. 1997
8. Maher Kevin. TEXTO. URL:ftp://ftp.funet.fi/pub/crypt/steganography/texto.tar.gz.
9. Wu Shufeng. Research on Information Hiding. Degree of master, University of Science and Technology of China, 2003.
10. Chang Chih-Chung and Lin Chih-Jen, LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm
11. Koehn Philipp. Europarl: A parallel corpus for statistical machine translation. In MT summit, 2005, vol. 5.
12. BNC database and word frequency lists. http://www.kilgarriff.co.uk/ bnc-readme.html.