http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 10 (5): 992-1000, 2011
ISSN 1812-5638 / DOI: 10.3923/itj.2011.992.1000

A Secure Steganographic Method via Multiple Choice Questions

[1]Lingyun Xiang, [1,2]Xingming Sun, [1]Yuling Liu and [3]Hengfu Yang
[1]College of Information Science and Engineering, Hunan University, Changsha, 410082, China
[2]College of Computer and Software, Nanjing University of Information Science and Technology,
Nanjing, 210044, China
[3]Department of Information Science and Engineering, Hunan First Normal University,
Changsha, 410205, China

Abstract: In this study, a novel secure steganographic method was proposed by taking advantage of Multiple-Choice Questions (MCQs). Selecting a series of MCQs to automatically generate a stego text could conceal secret information. The proposed method could achieve a considerable embedding bit rate while being able to hide more information by reordering the options of the generated MCQs. The final outputting stego texts were kept the same linguistic and statistic characteristics as normal texts, thus the hidden information had good imperceptibility and could survive under potential steganalysis attacks, which was also demonstrated by experiments. More experimental results showed that the average embedding bit rate achieved more than 1 bit per sentence, which was superior to the bit rate of most existing linguistic steganographic methods.

Key words: Steganography, linguistic steganography, steganalysis, embedding bit rate, multiple choice questions

INTRODUCTION

Steganography aims to hide the very presence of covert communications by imperceptibly embedding secret information into innocent-looking carriers, such as digital images, videos, texts, etc. With the increasing use of computer and the development of Internet, text is becoming increasingly popular in daily life. This motivates the development of linguistic steganography, which refers to employing the knowledge of natural language processing for camouflaging information into texts (Atallah *et al*., 2000).

Linguistic steganography has formed a new research field in steganography, which has recently attracted the attention of many researchers (Bergmair, 2007). The contemporary linguistic steganography in the literatures can be grouped into three categories in terms of the way of generating the stego texts.

The first class is generalized to the approach that directly produces a new natural-like stego text without the support of a given cover text. The primary methods mainly based on mimicking technique to convert the secret information into a text that obeys the statistical properties of another particular normal natural language text. In order to improve the readability and quality of the produced texts, probabilistic context-free grammars are combined with mimicking technique to generate stego texts (Wayner, 2002). Despite the fact that the generated texts have almost regular grammar and syntax, their contents are always meaningless and implausible to human readers. The transmission of these texts between different communication parties may raise suspicion by people inspection. Moreover, the existence of communicated secret information can be detected by using statistical characteristics of correlations between the general service words (Chen *et al*., 2008).

The second class which converts the cover text into a complete different normal-like stego text, can be considered as an extension to the first class mentioned earlier. Combining text-mimicking technique with PPT documents, Liu *et al*. (2008) converted secret information into innocuous sentences by exploiting the resource from a cover PPT document. Then the generated sentences are written into the note pages of the cover document to enhance the security. The efficiency and security of this steganography is greatly improved compared with the methods of the first class. Another typical method of this class is based on machine translation (Grothoff *et al*., 2005; Stutsman *et al*., 2006; Grothoff *et al*., 2009). Its critical point is to create multiple translations for each sentence in cover text and then select one of these translated sentences to encode

Corresponding Author: Xingming Sun, College of Information Science and Engineering, Hunan University,
No.252, Lushan South Road, Yuelu District, Changsha, 410082, China
Tel: 86-731-88821341 Fax: 86-731-88821780

information. The generated stego texts with the help of basic information provided by cover texts have better readability and quality than those produced by steganography of the first class. However, they also have a disadvantage of causing suspicious from text steganalysis (Meng *et al*., 2010). Steganalysis is the art of discovering secret information in digital media. Once a steganographic method is defeated by steganalysis, the secret information is no longer secure.

The approaches of slightly modifying an existing text to obtain a stego text are defined as the third class. This class of methods manipulates lexical, syntactic or semantic properties of a given text while preserving the meaning as much as possible to embed information (Topkara *et al*., 2005). Previous research on linguistic steganography is mainly focused on the approaches employing synonym substitution and syntactic transformation for hiding information.

The synonym substitutions-based approach embeds information in the way of replacing the word by its synonym to represent designated value. In theory, the modifications preserve the meaning of cover texts. However, simple synonym substitutions may cause potential wrong syntax and semantics since the synonyms with the identical or similar meanings still have semantic and pragmatic differences. Some more superior methods were proposed to overcome the vulnerability of synonym substitutions (Bolshakov, 2004; Topkara *et al*., 2006a). Bolshakov (2004) proposed to directly substitute an absolute synonym for a word and previously test a relative synonym whether it is semantic compatible with collocations containing the replaced word. Topkara *et al*. (2006a) used a quantitative resilience criterion to just choose one most appropriate alternative for every being replaced word in a text. However, the above efforts were just focused on improving the compatibility of the synonym substitutions with the context and do not consider the changes of the inherent statistical characteristics of synonyms sequence, such as occurrence frequency of the synonyms. The synonym substitution-based steganography has low resistance of the existing text steganalysis attack (Taskiran *et al*., 2006; Yu *et al*., 2008).

Another extensively concerned steganographic approach of the third category is based on syntactic transformations (Atallah *et al*., 2001; Liu *et al*., 2005; Wang *et al*., 2008; Topkara *et al*., 2006b; Murphy and Vogel, 2007; Meral *et al*., 2007; Meral *et al*., 2009; Kim, 2008). This approach expresses a sentence into a different semantically equivalent sentence structure for hiding information. For instance, it transforms an active sentence into a passive one without any serious meaning and grammaticality change. Atallah *et al*. (2001) were the first ones to employ transformations described in post-Chomskian generative syntax to design a natural language watermarking scheme. Topkara *et al*. (2006b) attempted to resolve the implementation challenges syntactic transformations-based steganography and designed an objective evaluation metric to evaluate the quality of the stego text. By taking the characteristics of the language into consideration, several sophisticated steganographic methods have been proposed using relevant particular syntactic transformations of different languages such as Chinese (Liu *et al*., 2005; Wang *et al*., 2008), Turkish (Meral *et al*., 2007; Meral *et al*., 2009), Korean (Kim, 2008), etc. Comparing with synonym substitution-based method, the syntactic transformation-based method has high robustness to resist attacks but it also demands deep structure analysis and its embedding capacity is limited. In practical, a perfect steganography tool is hard to be implemented and widely used for lacking enough natural language knowledge and matured reliable tools of natural language processing.

Linguistic steganography is gradually improved recently but much more efforts are need to avoid linguistic flaws in stego texts. However, it is a difficult task of building a complex natural language processing systems. Some linguistic steganographic methods possess of low popularization and application value or the stego texts in practical always have degradation of quality in theory. And texts always provide few redundancies of linguistics for embedding information leading to low embedding bit rate of linguistic steganography. Moreover, most steganographic methods have been defeated by using the distortions of statistics with the development of text steganalysis (Taskiran *et al*., 2006; Chen *et al*., 2008; Meng *et al*., 2010; Yu *et al*., 2008). These vulnerabilities and concerns of the above linguistic steganography motivate to develop a linguistic lossless steganographic method with high resistance of steganalysis attacks and improved embedding bit rate as well.

Multiple choices are the most popular and frequently used form in educational testing, professional certification examination and many other sorts of tests and surveys. Employing MCQs to hide information would not arouse suspicious from eavesdroppers. Considering the characteristics of MCQs, a secure steganography is proposed by taking MCQs as the cover carriers in this study. Firstly, by further investigation on MCQs, two inherent attributes are introduced for steganography, because of their peculiar nature. Then each MCQ is assigned to two determinate attribute values based on the character lengths and the order of its options. Selecting MCQs from a prepared MCQs bank according to their first attribute values will generate a stego text. And modifying the second attribute values of MCQs via equivalent transformations would embed more information.

MCQS-BASED STEGANOGRAPHIC METHOD

Within today's society, various examinations or tests designed to measure people's knowledge, skills, aptitude, etc., in many different fields such as the education, professional certification, psychology, the military, etc. MCQs are the basis of a significant portion of a test. In a test that has items formatted as MCQs, a candidate is given a number of set answers for each question and the candidate must choose which answer or group of answers is correct. With the convenience brought by Internet, many websites offers abundance MCQs for personal research and private study. Especially, some special MCQs banks of many subjects are provided. Before attending a formal examination, some exercises are always done by using these materials. Communication via a text including MCQs retrieved from a question bank is commonplace, thus it is convenient to transmit secret information by using the MCQs as the cover mediums and could not arouse suspicious from eavesdroppers.

Attributes of MCQs: A MCQ always consists of a stem and a series of options. A test requires a test taker to choose all appropriate answers or only one answer from a list of options. Two vocabulary MCQs of primary English are provided in the following as examples:

MCQ examples:

- MCQ 1. The match was _____ because of the heavy rain.
  A. called in   B. called for   C. called off   D. called out

- MCQ 2. The badminton players were trained under a well-known _____.
  A. judge   B. referee   C. umpire   D. coach

It is worthwhile to note that the MCQs in a test always serve the same subject. In many formal standardized tests, a test question is often retrievable from a question bank. From the appearance point of view, the arrangement of MCQs in a test paper is always automatically made; the order of MCQs can be adjusted based on requirements. On the other hand, options of well-written MCQs are independent of one another and consistent in logic and grammar to the stem. Therefore, the order of options can also be arbitrarily changed. With these characteristics, MCQs can be successfully used to hide information. Before describing the proposed steganography, two attributes, namely Q-attribute and P-attribute are introduced respectively for each MCQ.

Given a MCQ t = $<a_1,a_2,...,a_n>$, where a_j denotes the jth option. Let $L(a_j)$ denote the character length of the option a_j. After rearranging the options in an increasing lexicographic order, the options are denoted as a new vector $<a'_1,a'_2,...,a'_n>$. According to the character lengths of the reordered options, a Q-attribute denoted as g(t) is defined and calculated by:

$$g(t) = \sum_{j=1}^{n} 2^{n-j} b_j \qquad (1)$$

where, $b_j = L(a'_j) \bmod 2$.

Apparently, g(t) is a random n-bit integer, whose value is in range of 0 to 2^n. For the example MCQ 1 t_1 = <called in, called for, called off, called out>, after rearranging the options in an increasing lexicographic order, a new vector of options <called for, called in, called off, called out>, is obtained, then, $L(a'_1) = 10$ $L(a'_2) = 9$, $L(a'_3) = 10$, $L(a'_4) = 10$, thus its Q-attribute value is 4. In the same way, Q-attribute value of the example MCQ 2 figured out at 14. MCQs' Q-attribute values would be approximately independent and identically distributed (i.i.d.) in a large MCQs bank, since the character length of options in a MCQ is random and all options are mutually independent. MCQs with appointed Q-attribute values can be successfully retrievable from a large MCQs bank for embedding information.

Based on above analysis, we can also find that the options of a MCQ can be arranged into a required layout. Each possible arrangement can be assign to a unique value. For a given MCQ, the arrangement of its options will map to a determined value. Given the MCQ t = $<a_1,a_2,...,a_n>$, its n different options have n! number of permutations. Arranging the n! permutations in a certain order (lexicographic order, for example), the position of $<a_1,a_2,...,a_n>$ locating at is defined as the MCQ's P-attribute, denoted as f(t). In this study, we define the position starts at 0. The P-attribute value is a integer in range of 0 to n!-1. Generally, the options in a MCQ are always randomly placed in practical tests. Thus, the options can be reordered to make the MCQ's P-attribute equal to a required value in a certain range without influencing the meaning of the MCQ and bringing any errors.

Next we illustrate the procedure of obtaining the P-attribute value of a MCQ. Set the n options of MCQ t = $<a_1,a_2,...,a_n>$ in increasing lexicographic order is $a'_1,a'_2,...,a'_n$. For convenience, we set n = 4 for example. The

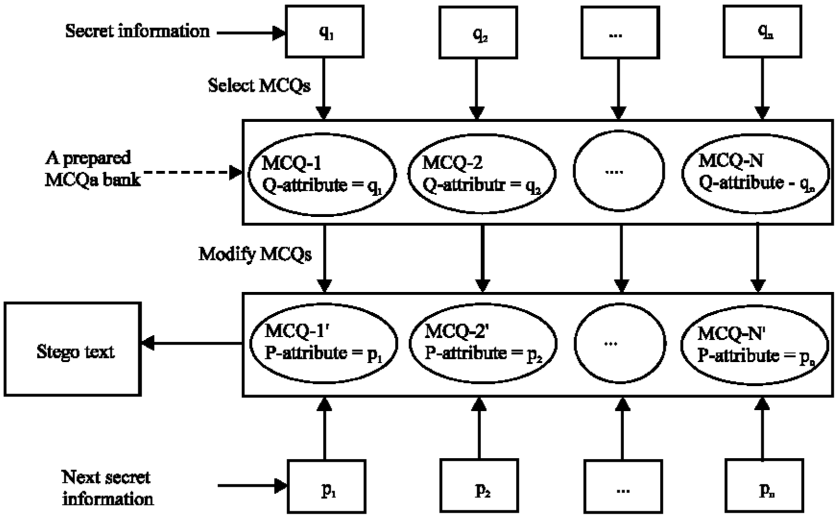
Secret information
q_1
q_2
...
q_n
Select MCQs
A prepared MCQa bank
MCQ-1 Q-attribute = q_1
MCQ-2 Q-attributr = q_2
....
MCQ-N Q-attribute - q_n
Modify MCQs
Stego text
MCQ-1' P-attribute = p_1
MCQ-2' P-attribute = p_2
...
MCQ-N' P-attribute = p_n
Next secret information
p_1
p_2
...
p_n

Fig. 1: Diagram of embedding process

24 permutations in increasing lexicographic order are listed as follows:

$$<a_1^{'},a_2^{'},a_3^{'},a_4^{'}>,<a_1^{'},a_2^{'},a_4^{'},a_3^{'}>,<a_1^{'},a_3^{'},a_2^{'},a_4^{'}>,<a_1^{'},a_3^{'},a_4^{'},a_2^{'}>,<a_1^{'},a_4^{'},a_2^{'},a_3^{'}>,$$
$$<a_1^{'},a_4^{'},a_3^{'},a_2^{'}>,<a_2^{'},a_1^{'},a_3^{'},a_4^{'}>,<a_2^{'},a_1^{'},a_4^{'},a_3^{'}>,<a_2^{'},a_3^{'},a_1^{'},a_4^{'}><a_2^{'},a_3^{'},a_4^{'},a_1^{'}>,$$
$$<a_2^{'},a_4^{'},a_1^{'},a_3^{'}>,<a_2^{'},a_4^{'},a_3^{'},a_1^{'}>,<a_3^{'},a_1^{'},a_2^{'},a_4^{'}>,<a_3^{'},a_1^{'},a_4^{'},a_2^{'}>,<a_3^{'},a_2^{'},a_1^{'},a_4^{'}>,$$
$$<a_3^{'},a_2^{'},a_4^{'},a_1^{'}>,<a_3^{'},a_4^{'},a_1^{'},a_2^{'}>,<a_3^{'},a_4^{'},a_2^{'},a_1^{'}>,<a_4^{'},a_1^{'},a_2^{'},a_3^{'}>,<a_4^{'},a_1^{'},a_3^{'},a_2^{'}>,$$
$$<a_4^{'},a_2^{'},a_1^{'},a_3^{'}>,<a_4^{'},a_2^{'},a_3^{'},a_1^{'}>,<a_4^{'},a_3^{'},a_1^{'},a_2^{'}>,<a_4^{'},a_3^{'},a_2^{'},a_1^{'}>$$

Each permutation has its own position in the ordered permutation set. From $<a'_1,a'_2, a'_3,a'_4>$ to $<a'_4,a'_3,a'_2a'_1>$, they are numbered 0 to n!-1, respectively. For a MCQ t = $<a_1,a_2,a_3,a_4>$, if it is happen to $a_1 = a'_1$, $a_2 = a'_2$, $a_3 = a'_3$, $a_4 = a'_4$, then its P-attribute f(t) = 0. If $a_1 = a'_1$, $a_2 = a'_3$, $a_3 = a'_2$, $a_4 = a'_4$, then f(t) = 2. For example MCQ 1 $t_1 = <a_1,a_2,a_3,a_4>$=<called in, called for, called off, called out>, after rearranging the options in an increasing lexicographic order is obtained $<a'_1,a'_2,a'_3,a'_4>$=<called for, called in, called off, called out>, then $a_1 = a'_2$, $a_2 = a'_1$, $a_3 = a'_3$, $a_4 = a'_4$, thus $f(t_1) = 6$. In the same way, the P-attribute of example MCQ 2 is calculated to 9.

The embedding process: With the help of the two attributes of MCQs, secret information can be securely embedded into MCQs. Before embedding secret information, we collected enormous MCQs to form different banks based on their subjects. In the embedding process, according to the secret information and the options number of a MCQ in the collected MCQs bank, two integer sequences will be calculated and encoded into MCQs. MCQs whose Q-attribute values orderly equal to the first integer sequence are selected to generate a stego text. Then, the options of the selected MCQs are reordered to make their P-attribute values accord with the second integer sequence. Finally, the generated stego text consisting of selected and modified MCQs is output. Diagram of embedding process is illustrated in Fig. 1.

Now, we describe the embedding process of our MCQs-based steganography in detail:

Step 1: Classify the MCQs in the prepared MCQs bank into s groups in terms of the number of their options. Record the options number of a MCQ in the kth group as n_k, k = 1,2,...,s. For each MCQ t, set a marker Mar(t) and make Mar(t) = 0 in default

Step 2: Encrypt the secret information using an encryption algorithm and convert the encrypted information into a larger integer denoted as M

Step 3: Initial i = 1, k = s

Step 4: Set n = n_k, calculate $M' = \lfloor M/2^n \rfloor$, q_i = M mod 2^n, in other words, q_i equals to the value of the least n bits of M

Step 5: Select a MCQ t = $<a_1,a_2,...,a_{ln}>$ whose Q-attribute g(t) = q_i and Mar(t) from the kth groups in the MCQs bank and modify Mar(t). If there are not any satisfied MCQ can be successfully found, then k = k-1 and return to step 4

Step 6: Let M = M', calculate $M' = \lfloor M/n! \rfloor$, p_i = M mod n!, M = M' and reorder the options of MCQ t into a new arrangement to make its P-attribute f(t) = p_i

Step 7: If M =0, finish the embedding process and output the stego text composed of the selected MCQs or else i = i+1 and return to step 4

Extraction process: In the extraction process, a larger integer is first obtained from the two attribute values of

MCQs in the stego text. Then translating the integer into decrypted bytes can retrieve the secret information. The Extraction process involves the following three steps:

Step 1: For each MCQ $t_i = <a_{i1},a_{i2},...,a_{in}>$ in the stego text, calculate its Q-attribute $g(t_i)$ and P-attribute $f(t_i)$, the number of options of t_i is denoted as n_i, $i = 1,...N$, N is the total number of MCQs

Step 2: Let M_{N+1}, repeatedly calculate

$$M_i = (M_{i+1} \times n_i! + f(t_i)) \times 2^{n_i} + g(t_i) \tag{2}$$

Step 3: Retrieve the embedding information by converting M_1 into bytes and decrypting the decoded information

RESULTS AND DISCUSSION

The proposed method can avoid making transformations under requirements of natural language processing techniques, such as natural language generation, understanding, part-of-speech tagging, syntax or semantic analysis, etc. Its implementation just needs a shallow analysis of identifying the components MCQs and a collection of enormous MCQs. Thus, it is easy to be implemented. In the experiments, we implemented it and carried out the performance evaluation.

First, many MCQs of different subjects such as English vocabulary, grammar, PMP (Project Management Professional) exam, Computer Fundamentals, IBM certification were collected from Internet. But, in practice, some MCQs are not well designed which include overlapping alternatives when one or more options are a subset of another. For example, the options contain the content of "all of the above", "none of the above", etc. These poorly constructed and ineffective options are unsuitable for our steganography. After removing the inappropriate MCQs, we obtained 9561 MCQs for present experiments.

Now, we provide an example to further explain the embedding process and extraction process of the proposed MCQs-based steganography. Assuming the secret information is "China", its ASCII value is 67 104 105 110 97. For simplicity, the secret information is directly hidden into MCQs without being encrypted. The larger integer converted from "China" is M = 289514548833. Suppose the options number of each MCQ in the prepared MCQs bank is 4, i.e., n = 4. According to the embedding process, we calculate $M' = \lfloor M/2^n \rfloor$, $q_1 = M \bmod 2^n$ to obtain $q_1 = 1$ and calculate $M = M'$, $M' = \lfloor M/n! \rfloor$, $p_1 = M \bmod n!$, $M = M'$ to obtain $p_1 = 14$. Then, we select a MCQ whose Q-attribute value is 1 and rearrange its options to make the P-attribute value equal to 14. In the same way, we figure out $q_2,q_3,q_4,q_5 = 9,4,9,13$, p_2,p_3,p_4, $p_5 = 4,0,7,0$ and select four different MCQs whose Q-attribute values are 9, 4, 9 and 13 in order and modify their options' order to make their P-attribute values equal to 4, 0, 7, 0. Finally, the selected MCQs are written into a stego text. Conversely, the information can be recovered by converting the integer, which is obtained from Eq. 2 with the Q-attribute and P-attribute values of MCQs in the stego text, into characters. A concrete stego text with the MCQs from an English vocabulary MCQs bank is given in the Appendix A.

Security analysis: Transmitting a stego text generated by our methods is much reasonable and secure. As the used MCQs are derived from existing test banks, stego texts consisting of MCQs are meaningful and comprehensible. In addition, the linguistic characteristics of both a single MCQ and whole MCQs are preserved, as the contents of MCQs are not modified during the embedding process. Thus, present method can obtain good imperceptibility.

During the embedding process, secret information is successfully hidden into MCQs by utilizing their Q-attributes and P-attributes. The two attribute values are determined by the length and arrangement of options. As we know, the least significant bit of the option' character length and arrangement of options in a MCQ is random. In a normal text composed of MCQs, the MCQs' Q-attribute and P-attribute values are approximately i.i.d. In the rest of this study, the normal text is especially the one just contains MCQs. On the other hand, the Q-attribute values of MCQs in a stego text corresponding to random integers from the secret information pieces are also i.i.d. In addition, the P-attribute values were maintained and evenly distributed, as the modifications on P-attribute values made by the proposed method are random according to the embedded information. In one word, the statistical characteristics of stego texts with MCQs are coincided with that of normal texts in terms of the distributions of Q-attribute and P-attribute. This will make steganalysis attackers fail to distinguish the generated stego texts from normal texts.

To demonstrate the above theory analysis, we conduct experiments on the distributions of Q-attribute and P-attribute in stego texts and normal texts. First, 50 stego texts containing 50-200 MCQs from four prepared MCQs banks were generated. The MCQs in a text is about the same subject. The subjects of the four banks are English vocabulary, PMP, Computer Fundamentals and IBM certification, respectively. The secret information embedded in each stego text is different and encrypted by using RC4 encryption algorithm. For comparison, 50 normal texts consisting of 50-200 number MCQs are

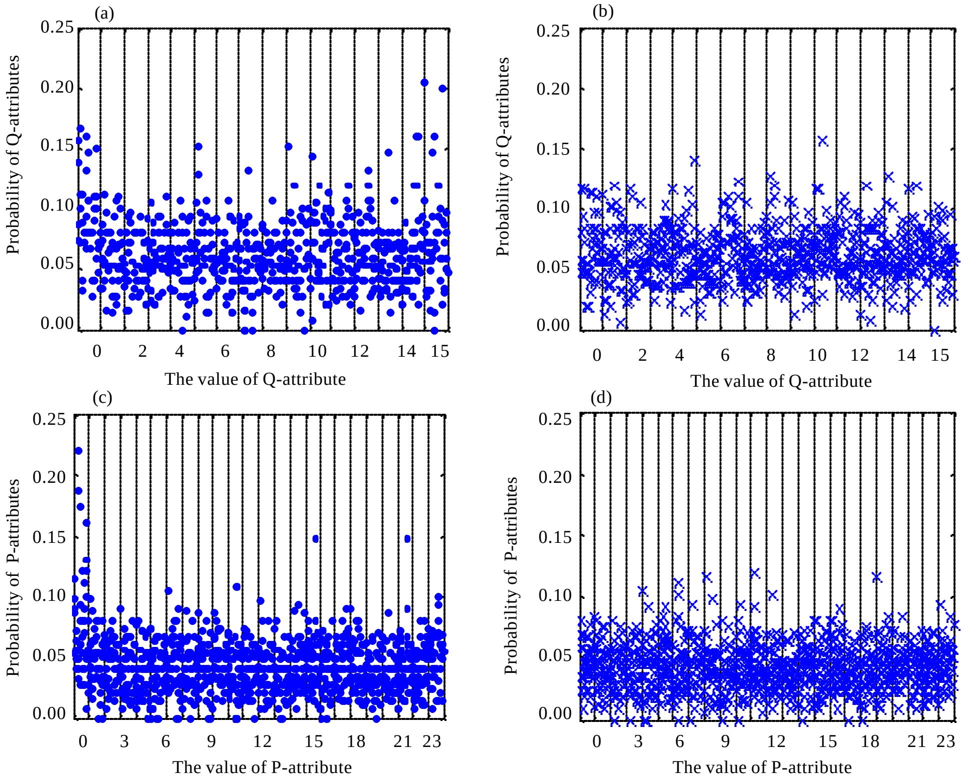

Fig. 2: The distribution of Q-attribute and P-attribute in normal and stego texts (a) The distribution of Q-attribute in normal texts (b) The distribution of Q-attribute in stego texts (c) The distribution of P-attribute in normal texts (d) The distribution of P-attribute in stego texts

prepared. Both the MCQs in normal texts and stego texts have 4 options. The Q-attribute and P-attribute of a MCQ has 2^4 and 4! possible values, respectively.

For each text, 16 probabilities of all the potential Q-attribute values can be obtained. Thus, 800 probabilities can be obtained from 50 sample normal texts. We order these 800 probabilities according to the ascending order of the Q-attribute values. And the probabilities of the same Q-attribute value in different sample normal texts are placed together. For example, the first 50 probabilities are the ones when the Q-attribute equals to 0 in 50 different sample texts while the last 50 probabilities are the ones when the Q-attribute equals to 15 in 50 different sample texts. All 800 probabilities are exhibited in Fig. 2a. In the same way, the probabilities of all the potential Q-attribute values in 50 sample stego texts are exhibited in Fig. 2b. The probabilities of all the potential P-attribute values in 50 sample normal texts and 50 stego texts are exhibited in Fig. 2c and d, respectively.

From Fig. 2a-d, it can be observed that both of the distributions of Q-attribute and P-attribute in stego texts are approximate to those in normal texts. It is much difficult of taking the probability of Q-attribute and P-attribute as clue to detect the existence of secret information hidden in MCQs. Consequently; the proposed method has high resistance of steganalysis attack and can provide high-level security for secret information.

Embedding bit rate: In this study, embedding bit rate is quantified in terms of embedded bits per language unit (sentence, for example). There are three parameters which affect the embedding bit rate. The three parameters are the number of used MCQs in a stego text, the average number of sentences per MCQ and the options number in a MCQ.

For a stego text composed of N MCQs, each with n options, let the average sentences per MCQ is L, then embedding capacity of the stego text is given by :

$$C=nN+\left\lfloor \log_2(n!)^N \right\rfloor \text{bits} \qquad (3)$$

The embedding bit rate is defined as:

Table 1: Average embedding bit rate comparison of stego texts of different subjects

Subject of the stego texts	Average sentences per MCQ	Average embedding bit rate (bit/sentence)
English vocabulary	5.45	1.53
PMP exam	5.81	1.45
Computer Fundamentals	5.04	1.68
IBM certification	7.36	1.15

$$r = \frac{C}{N \times L} \text{ bit / sentence} \qquad (4)$$

It is obvious that with N increasing, the embedding bit rate r increases. After substituting Eq. 3 into Eq. 4, the transformation of Eq. 4 is:

$$r = \frac{nN + \lfloor \log_2 (n!)^N \rfloor}{N \times L} \qquad (5)$$

The smaller L is and the larger n is, the larger r is. In theory, $r \geq {}^{(n + \lfloor \log_2 n! \rfloor)}\!/_{L}$. The minimum of L is $L_{min} = 1+n$, when all the stem of each used MCQ and its options just contain one sentence. The options consisting of fragmentary sentence such as a word or a phrase are all treated as a sentence. Thus the maximum embedding bit rate:

$$r_{max} \geq \frac{n + \lfloor \log_2 n! \rfloor}{1 + n}$$

when n = 3,4,5, $r_{max} \geq 1.396$, 1.717, 1.984. For a stego text with large numbers of MCQs and at least 5 options in each MCQ, its embedding bit rate is possible to exceed 2 bit/sentence.

In experiments, we calculated the embedding bit rate of above-mentioned 50 stego texts. The MCQs including figures were removed and not used for embedding information. The practical average embedding bit rate of stego texts with MCQs from different subjects are listed in Table 1. The experimental results show that it is easy to achieve an embedding bit rate of more than 1 bit/sentence using the proposed method which is superior to results of most linguistic steganographic methods.

After investigating the previous linguistic steganographic methods, we can find that their embedding bit rates attained from the literatures are almost always less than one. The machine translation-based method proposed in Stutsman *et al.* (2006) has just an embedding bit rate of 0.33 bit/sentence, as more embedding space must be provided for concealing the extra head information which is used for recovering the hidden information without the source text. For the synonym substitutions-based method, the word should be substituted with a synonym in the same sense to preserve the meaning of the sentence. However, most words always have multiple senses and the synonyms under different senses of the same word are always not agreed. Each synonym substitution must be tested using natural language processing knowledge such as word sense disambiguation to determine whether the substitution is suitable or not. Thus the number of successful synonym substitutions for hiding information is relatively small leading to a low embedding bit rate. For example, Topkara *et al.* (2006a) approximately achieved an embedding bit rate of 0.67 bit/sentence. The embedding bit rate of syntactic transformations-based methods are often very low, as the available syntactic transformations are limited and not all transformations can be applied to each sentence in a text. The scheme reported in Atallah *et al.* (2001), yielded a bit rate of 0.52 bit/sentence based on the demonstration sample developed by the authors. And referring the method in Murphy and Vogel (2007), just a bit rate of 0.3 bit/sentence is claimed by the authors. The bit rate in Wang *et al.* (2008) is calculated to be approximately 0.35 bit/sentence using Chinese syntactic transformations based on the data in this literature. Finally, Meral *et al.* (2009) used a especially suitable language Turkish, which has more syntactic transformations to conceal information than English language and provided a higher embedding bit rate of 0.81 bit/sentence. Therefore, compared the abovementioned methods with ours, it is obvious that our method can be more effective for covert communication with larger embedding bit rate.

CONCLUSION

This study is the first attempt at constructing covert communication based on MCQs. The proposed method can transmit secret information by automatically generating new texts with MCQs, meanwhile conducting modifications on the options' order of MCQs for embedding more information. Regarding the imperceptibility, no linguistics distortions or errors are brought into a stego text generated by the proposed method, making the secret information greatly imperceptible. In addition, the statistical characteristics of MCQs are well preserved which have been proven by theoretical analysis and experimental results. Consequently, the existence of secret information would not be detected by potential steganalysis. The experimental results have also shown that the proposed method achieved higher embedding bit rate than most previous linguistic steganographic methods. In conclusion, the proposed MCQs-based steganography has the advantages of good imperceptibility, high resistance of steganalysis attacks, large embedding bit rate and ease of implementation. It is a practical, effective and secure approach for covert communication.

ACKNOWLEDGMENT

This study was supported in part by National Natural Science Foundation of China (No. 60736016, No. 60973128, No. 60973113, No. 61073191 and No. 61070196), in part by National Basic Research Program of China (No. 2009CB326202, 2010CB334706) and in part by Hunan Provinicial Natural Science Foundation of China (No. 10JJ4041).

APPENDIX A

A.1. An example of the Stego text with the secret information "China".

English vocabulary exercise
Pick the correct answer:

1. I find the other ________ of "The Frog Prince" more entertaining. The original story is too predictable.
A. type B. tale C. kind D. version
2. After the hurricane, the government decided to spend millions of dollars to ________ hundreds of damaged buildings.
A. build B. restore C. decorate D. repair
3. The match was ________ because of the heavy rain.
A. called for B. called in C. called off D. called out
4. As Rui Long liked collecting erasers; his friends gave him the ________ "Rubber Boy".
A. greeting B. brand C. surname D. nickname
5. A ________ of lions is roaming in the nearby forests.
A. flock B. pride C. school D. swarm

REFERENCES

Atallah, M.J., C.J. McDonough, V. Raskin and S. Nirenburg, 2000. Natural language processing for information assurance and security: An overview and implementations. Proceedings of the 2000 Workshop on new Security Paradigms, Sept. 10-13, ACM Press, New York, USA., pp: 51-65.

Atallah, M.J., V. Raskin, M. Crogan, C. Hempelmann, F. Kerschbaum, D. Mohamed and S. Naik, 2001. Natural language watermarking: Design, analysis and a proof-of-concept implementation. Proceedings of the 4th International Workshop on Information Hiding, April 25-27, Springer, pp: 185-199.

Bergmair, R., 2007. A comprehensive bibliography of linguistic steganography. Proceedings of the 9th Conference on Security, Steganigraphy and Watermarking of Multimedia Contents, Feb., 2007, 6505: 65050 W-1-65050W-6

Bolshakov, I. A., 2004. A method of linguistic steganography based on collocationally-verified synonymy. Proc. Int. Workshop Inform. Hiding, 3200: 180-191.

Chen, Z., L. Huang, Z. Yu, W. Yang, L. Li, X. Zheng and X. Zhao, 2008. Linguistic steganography detection using statistical characteristics of correlations between words. Proc. Int. Workshop Inform. Hiding, 5284: 224-235.

Grothoff, C., K. Grothoff, L. Alkhutova, R. Stutsman and M. Atallah, 2005. Translation-based steganography. Proc. Int. Workshop Inform. Hiding, 3727: 219-233.

Grothoff, C., K. Grothoff, R. Stutsman, L. Alkhutova and M. Atallah, 2009. Translation-based steganography. J. Comput. Secur., 17: 269-303.

Kim, M.Y., 2008. Natural language watermarking for Korean using adverbial displacement. Proceedings of the International Conference on Multimedia and Ubiquitous Engineering, April 24-26, Busan, Korea, pp: 576-581.

Liu, Y., X. Sun and Y. Wu, 2005. A natural language watermarking based on Chinese syntax. Adv. Nat. Comput., 3612: 958-961.

Liu, Y., X. Sun, Y. Liu and C.T. Li, 2008. MIMIC-PPT: Mimicking-based steganography for microsoft power point document. Inform. Technol. J., 7: 654-660.

Meng, P., L. Hang, Z. Chen, Y. Hu and W. Yang, 2010. STBS: A statistical algorithm for steganalysis of translation-based steganography. Proc. Int. Workshop Inform. Hiding, 6387: 208-220.

Meral, H.M., E. Sevinc, E. Unkar, B. Sankur, A.S. Ozsoy and T. Gungor, 2007. Syntactic tools for text watermarking. Proc. SPIE Int. Conf. Secur. Steganography Watermarking Multimedia Contents, 6505: 65050X-1-65050X-12.

Meral, H.M., B. Sankur, A.S. Ozsoy, T. Gungor and E. Sevinc, 2009. Natural language watermarking via morphosyntactic alterations. Comput. Speech Lang., 23: 107-125.

Murphy, B. and C. Vogel, 2007. The syntax of concealment: Reliable methods for plain text information hiding. Proc. SPIE Conf. Security Steganography Watermarking Multimedia Contents, 6505: 65050Y-1-65050Y-12.

Stutsman, R., M. Atallab, C. Grothoff and K. Grothoff, 2006. Lost in just the translation. Proceedings of the 21st Annual ACM Symposium on Applied Computing, April 23-27, Association for Computing Machinery, Dijon, France, pp: 338-345.

Taskiran, C.M., U. Topkara, M. Topkara and E.J. Delp, 2006. Attacks on lexical natural language steganography systems. Proc. SPIE Conf. Security Steganography Watermarking Multimedia Contents, 6072: 97-105.

Topkara, M., C.M. Taskiran and E.J. Delp, 2005. Natural language watermarking. Proc. Int. Conf. Secur. Steganogr. Watermarking Multimedia Contents, 5681: 441-452.

Topkara, U., M. Topkara and M.J. Atallah, 2006a. The hiding virtues of ambiguity: Quantifiably resilient watermarking of natural language text through synonym substitutions. Proceedings of the 8th Workshop on Multimedia and Security, Sept. 26-27, ACM Press, Geneva, Switzerland, pp: 164-174.

Topkara, M., G. Riccardi, H. Dilek and M.J. Atallah, 2006b. Natural language watermarking: Challenges in building a practical system. Proc. SPIE, 6072: 106-117.

Wang, X., X. Sun, Y. Liu and Y. Liu, 2008. Natural language watermarking using chinese syntactic transformations. Inform. Technol. J., 7: 904-910.

Wayner, P., 2002. Disappearing Cryptography: Information Hiding: Steganography and Watermarking. 2nd Edn., Morgan-Kaufmann, San Mateo, CA., pp: 81-128.

Yu, Z., L. Huang, Z. Chen, L. Li, X. Zhao and Y. Zhu, 2008. Detection of synonym-substitution modified articles using context information. Proceedings of the 2nd International Conference on Future Generation Communication and Networking, Dec. 13-15, IEEE Computer Society, Hainan, China, pp: 134-139.