A Natural Language Steganography Technique for Text Hiding Using LSB's

Dr. Hana'a M. Salman

Received on:3/7/2007
Accepted on:5/12/2007

Abstract

Steganography is the art of hiding, and transmitting information using apparently innocent carrier without expose any suspicion. This paper present a natural language steganography technique, which is different from all the natural language steganography technique, that uses structure of the sentence constituents in natural language text in order to insert a secret hidden information, or all others techniques, which hide information by modifying the appearance of text elements, such as lines, words, or characters. The proposal technique use the secret hidden text information to generated the stego-cover carrier text by using algorithms depend on natural language processing, particularly text generation field. A survey for natural language terminology, techniques, and tools for text processing, a natural language steganography technique and its difficulties to implement methods like LSB's is presented. The results shows that, the proposal technique is, a successful one in implementing methods like, LSB's for natural language bit steganography .

Keywords: natural language steganography, natural language processing, secret text hiding, LSB's, text generation field

الخلاصة:

الكتابة المغطاة تعرف على أنها فن أخفاء المعلومات المرسلة، في نوا قل بريئة المظهر وبدون إثارة الشكوك. هذه المقالة تقدم تقنية لإخفاء في نص مكتوب باللغة الطبيعية. أن الطريقة المقترحة هي مختلفة عن الطرق السابقة لإخفاء نص في نص مكتوب بأحد اللغات الطبيعية والذي يستعمل فيه الهيكل الأساسي لبناء الجمل باللغة الطبيعية. وكذلك الطرق التي تعتمد على تغير مظهر مكونات النص كوسيلة لإخفاء المعلومات كالخطوط، للكلمات، أو الأحرف. أن التقنية الإخفاء المقترحة تستعمل النص السري المطلوب إخفائه لتوليد الغطاء الحامل للمعلومات السرية باستعمال خوارزميات تعتمد على تقنيات معالجة اللغة الطبيعة وبصورة خاصة حقل توليد النصوص. نظرة عامة عن مصطلحات وتقنيات و سبل وأساليب معالجة النصوص المكتوبة باللغة الطبيعية. تقنية للإخفاء في نص مكتوب باللغة الطبيعية والمعوقات التي تواجه طريقة الجزء الأقل أهمية. الطريقة المقترحة تظهر أمكانية استعمال الجزء الأقل أهمية كتقنية لإخفاء في نص مكتوب باللغة الطبيعية.

*Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq

1. Introduction

Cryptography means secret writing. A closely related area to cryptography is Steganography, which literally means covered writing as derived from Greek and deals with the hiding of messages so that the potential monitors do not even know that a message is being sent. It is different form cryptography where they know that a secret message is being set. Figure (1) shows a schematic diagram of a typical steganography system. Generally, the sender performs the following operations [1]:

1. Write a non-secret cover-message.
2. Produce a stego-message by concealing a secret embedded message on the cover-message by using a stego-key.
3. Send the stego-message over the insecure channel to the receiver.
4. At the other end, on receiving the stego-message, the intended receiver extracts the secret embedded message from the stego-message by using a pre agreed stego-key.

Historical tricks include invisible inks, tiny pin punctures on selected characters minute differences between handwritten characters, etc. Note that the procedures of message concealing and message extracting in steganoraphic are more or less the same as the message encryption and message decryption in cryptography. It is this reason that steganography is often used together with cryptography. In general steganography system can either be secret or public. In public–key steganoraphic system different keys are used for message concealing "embedding" and message extraction [1].

Natural language steganography techniques, which aims to hide secret information in text documents by manipulating the semantic and/or syntactic structure of sentences. Others techniques which alter the appearance of text elements, such as lines, words, or characters as descript in [2].

The objective of this paper is to proposal a natural language steganography techniq Stego-Key ıg secret text using LSB's. The proposal technique is different from all the previous methods in that cover carrier text is generated and

*Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq

translated into other language at final step. The implementation is made by four main steps and these are: the first step is random number generation (a large prime number is used to generate the secret key). The second step is steganography method implementation (The bits of secret message is embedded in the LSB's of the result secret key). The third step is text generation (the results stego-carrier bytes is used to generate a text). The last step text translation (the result text is translated form English into Arabic language).

The organization of the paper is as follows: In Section 2, Basic concepts of natural language processing techniques and available resources, which can be employed to develop natural language steganography systems are introduced. The unique difficulties in natural language steganography caused by the structure of language are introduced in Section 3. Surveys of current state of the art in natural language steganography 4. In Section 5 the proposal technique with a test, followed by a summary and conclusions in Section 6.

2. Natural Language Processing

Natural Language Processing (NLP) aims to design algorithms that will analyze, understand, and generate natural language automatically. In bellow subsection, a briefly introduction to NLP techniques, and resources that are of interest for information hiding in natural language text [2]

2.1. Data Resources

Success of an information hiding system depends on obtaining good models of the cover medium which can only be achieved with large data sets. A statistically representative sample of natural language text is referred to as a corpus. Since most of NLP research is based on statistical analysis and machine learning systems, large corpora in machine readable form are essential. Therefore, a number of corpora in electronic form have been created and are commonly used in NLP research. In order to make the corpora more useful for NLP research, they are usually annotated with extra information. In addition to corpora, there are also electronic

*Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq

dictionaries available that are designed as large databases of lexical relations between words [2].

2.2. Linguistic Transformations

In order to embed information in natural language text a systematic method for modifying, or transforming, text is needed. These transformations should preserve the grammaticality of the sentences. Ideally, also require is that the differences in sentence meaning caused by the transformations should not be noticeable[2].

Generally three types of transformations are used for modification [2]: *Synonym substitution*: Synonym substitution has to take the sense of the word into consideration. In order to preserve the meaning of the sentence the word should be substituted with a synonym in the same sense. *Syntactic transformations*, such as passivization and clefting, which change the syntactic structure of a sentence with little effect on its meaning. *Semantic transformations*: a method to generate meaning-preserving semantic transformations is by using noun phrase coreferences. Two noun phrases are coreferent if they refer to the same entity.

2.3. Natural Language Parsing

In NLP parsing is defined as processing input sentences and producing some sort of structure for them. The output of the parsing may either be the morphological, syntactical, or semantically structure of the sentence or it may be a combination of these. Parsing is essential to get more information about the sentence structure and the roles of the constituent words in this structure [2].

2.4. Natural Language Generation

The natural language generation "NLG" task is defined as the process of constructing natural language output from non-linguistic information representations according to some communication specifications. The components of a typical NLG system are illustrated in Figure (2). There are several fully implemented NLG systems freely available for research purposes [2].

*Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq

2.5. Text Paraphrasing

The task of text paraphrasing entails changing text parameters such as length, readability, and style for a specific purpose without losing the core meaning of the text. Therefore, text paraphrasing is directly related to natural language steganography. Text paraphrasing is also similar to machine translation; however, rather than converting text from one language to another, it is modified from one form to another within the same language. Paraphrasing systems are mainly based on creating or collecting sets or pairs of semantically equivalent words, phrases, and patterns [2].

3. Difficulties in Natural Language Steganography

The goals of steganography in natural language are, to hide secret information by embedded it into the cover-carrier in a discreet manner, such that the alterations are imperceptible when the stego-carrier data is obsessive. On the other hand, language has a discrete and syntactical nature that makes such techniques more difficult to apply. Specifically, language, and as a result its text representation, has two important properties that [2]:

1. Sentences have a combinatorial syntax and semantics. That is, structurally complex (molecular) representations are systematically constructed using structurally simple (atomic) constituents, and the semantic content of a sentence is a function of the semantic content of its atomic constituents together with its syntactic/formal structure.
2. The operations on sentences are causally sensitive to the syntactic/formal structure of representations defined by this combinatorial syntax.

The atomic/syntactical nature of language brings about unique challenges for in natural language steganography. For example, deriving an analog of least significant bit (LSB's) embedding used in image steganography that alters text locally,

*Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq

i.e., based on words, without making perceptually significant changes to sentence structure is a hard problem. This is due to the fact that even small local changes in a sentence can change its semantics and/or make it ungrammatical. The only current local alterations techniques used are the synonym substitution methods in natural language steganography [2].

4.Natural Language Steganography Techniques

Work in natural language steganography is depends on the NLP techniques and tools. A review to the previous work done in NL steganography is presented in bellow subsections.

4.1. Using Probabilistic Context-Free Grammars to Generate Cover Text

A probabilistic context-free grammar "PCFG" is a commonly used language model where each transformation rule of a context-free grammar has a probability associated with it. A PCFG can be used to generate strings by starting with the root node and recursively rewriting it using randomly chosen rules. Conversely, a string belonging to the language produced by a PCFG can be parsed to reveal the sequence of possible rules that produced it. The problem with this method is that even within limited linguistic domains, deriving a PCFG that models natural language is a daunting task. Also, some aspects of language cannot be modeled by context-free grammars at all. Because of these reasons cover texts produced by PCFGs tend to be ungrammatical and nonsensical. This makes it very easy for native speakers to detect such texts, which defeats the steganoraphic purpose of the method. Therefore, this method can only be used in communication channels where computers act as wardens [2].

4.2. Information Embedding Through Synonym Substitutions

The simplest method of modifying text for embedding of a payload is to replace selected words by their synonyms so that the truth values of the modified sentences are preserved, as Wordnet, may be used to find, for

*Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq

each selected word w in text, the synonym set, which is defined as a set of words that are synonymous with w. words in a synonym set are indexed according to their alphabetical order. During embedding the selection process picks a subset of words from the text for replacement [2].

4.3. Generating Cover Text Using Hybrid Techniques

The NICETEXT system for the generation of natural-like cover text according to a given payload uses a mixture of both of the methods discussed above. The system has two components: a dictionary table and a style template. The dictionary table is a large list of (type, word) pairs where the type may be based on the part-of-speech of word or its synonym set. Such tables may be generated using a part-of-speech tagger or Wordnet. The dictionary is used to randomly generate sequences of words. The style template, which is conceptually similar to the PCFG, improves the quality of the cover text by selecting natural sequences of parts-of-speech while controlling the word generation, capitalization, punctuation, and white space [2].

5. The proposal Natural Language Steganography Technique

The aim of steganography methods is hide secret information via a apparently innocent carrier without expose any suspicion [3]. For a secure communications, many steganography application methods try to generate the digital cover carrier, and then used it for secret information hiding [4]. Since the used cover carrier type is natural language, and by takes in consideration the difficulties of using natural language steganography, which related to the structure of the language. The proposal steganography technique, embedded each bit of the secret hidden text bits, with LSB's of random number generator, seeded by a prime number. The result letters is converted into meteorological words. Takes raw meteorological data, and the FOG to generate the stego-carrier cover text and send it after translated into the Arabic language via the internet.

*Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq

Many NLG, and translator systems are freely available for research purposes.

In this section, the proposal natural language steganography technique for secret text data hiding algorithm is introduced in subsection 5.1, and a test example for the proposal technique is introduced in subsection 5.2.

5.1 Algorithm for the Proposal Natural Language Technique

With idea of using LSB's steganography technique, to hide secret text information into the natural language, using random number generator, and NLG for generating the stego-carrier, and finally a means of translation form one language into the other as presented in Figure (3). The implementation to the proposal technique is made by using FOG, and translator systems. While the result program is implemented using matlab ver7 programming language the algorithms of random number generator [1] ,hiding and Extracting is restricted bellow sub sections.

5.1.1 Random Number Generator

Input: (p) prime number, k: bits number of secret text information , period length $l = n$,a,b

Output: ($r_0, r_1, ..., r_{k-1}$)

Process:

Step1: Initialization: input r_0, a, b, n and k , $j \leftarrow 1$

Step2: Compute $r_j \leftarrow (ar_{j-1} + b)(\mathrm{mod}\, n)$

Step3: $r_i = r_i (\mathrm{mod} \mathit{number of letter in the used language})$,and print r_j

Step 4: increase j: $j \leftarrow j + 1$. If $j \geq k$, then go to step 5, else go to step 2

Step 5: End

5.1.2 Proposal Hiding Algorithm

Input: Secret text information

Output: Stego-cover carrier text written in any language.

Process:

Step1: Convert the input secret text information into bits ($mbit_i$).

Step2: Use a random number generator which initialized by a

*Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq

prime number (p) to generate (k) numbers ($r_0, r_1, \ldots, r_{k-1}$)

Step3: n_i is the number result from Hiding ($mbit_i$) into the LSB's of (r_i)

Step4: w_i is the letter result of Char (n_i)

Step5: gfw_i is the Generate forecast word beginning with letter w_i .

Step6: Repeat step (1-3) until the end of ($mbit_i$).

Step7: Use FOG, and gfw_i to generate the bilingual text in English / French text.

Step8: Translate the results form step 4, into an Arabic language "which represents the stego-cover carrier text".

Step9: End

5.1.2 Proposal Extracting Algorithm

Input: Stego-cover carrier text.

Output: Secret text information

Process:

Step1: Translate back the stego-cover carrier text to English / French

Step2: Extract the meteorological words (gfw_i)

Step3: Extract the first letter (w_i).

Step3: n_i is the ASCII value of (w_i)

Step4: Extract the LSB's for each (n_i).

Step5: Convert each eight ($mbit_i$) into a letters "which represents the secret text information"

Step6: End

5.2 Test Example

Suppose the secret text information is {big day}. Use the above algorithms for the proposal steganography technique. In bellow subsection is the results of implementations.

5.2.1 Proposal hiding implementation

Input: {big day},

*Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq

Output: تكسر الجليد في انهر إنكلترا. على تهب الرياح الاستوائية نحو ثقيل من الجنوب إلى الشمال،

Process:

1: Convert the input secret text information into bits ($mbit_i$).

{Big day} → {1000010,1101001,1100111,0100000,1000100,1100001,1111001}

2: Use a random number generator which initialized by a prime number (5) to generate (49) random numbers ($r_0, r_1, ..., r_{k-1}$) {r0=5, a=11, b=73, n=1399, k=56} → {128,82,975,1005,1335,768,127,71,854,1073,684,602,1099,970,950,730,1108,1069,640,118,1371,1164,286,421,507,54,667,415,441,727,1075,706,844,963,873,1282,185,709,877,1326,669,437,683,591,978,1038,299,564,681,569,736,1174,396,232,1226,968}

3: n_i is the number result from Hiding ($mbit_i$) into the LSB's of ($r_i \bmod 26$)

($r_i \bmod 26$) → {3, 6, 14,2,20,18,21,4,17,10,8,11,19,7,0,23,1,9,22,15,13,16,24,12,5,3,6,14,2,20,18,21,4,17,10,8,11,19,7,0,23,1,9,22,15,13,16,24,12,5,3,6,14,2,20,18}

{ni} → {3,7,15,2,21,19,21,4,17,11,9,10,19,7,0,22,0,8,22,14,13,16,25,12,4,2,6,15,2,20,19,20,5,16,10,9,10,19,6,0,22,0,8,23,15,12,16,25,13}

4: w_i is the letter result of Char (n_i)

{w_i} → {D,H,P,C,V,T,E,R,L,J,K,T,H,A,W,A,I,W,Q,M,Q,Z,M,Z,C,G,P,C,V,T,V,F,Q,B,A,J,K,T,G,Q,W,A,I,X,P,M,Q,Z,N}

5: gfw_i is the Generate forecast word beginning with letter w_i.

{gfw_i} → {debacle, heavy,…}

6: Use FOG, and gfw_i to generate the bilingual text in English / French text.

{Debacle is take place in all rivers in England , heavy wind from south to north.,…}

7: Translate into an Arabic language:

*Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq

تكسر الجليد في انهر إنكلترا. على نحو }
ثقيل تهب الرياح الاستوائية من الجنوب إلى
{الشمال.،.....

8: End

5.2.2 Proposal Extracting Algorithm

Input: { تكسر الجليد في انهر إنكلترا.
على نحو ثقيل تهب الرياح
الاستوائية من الجنوب إلى
{الشمال.،.....

Output: {big day}

Process:

1: Translate back:

{Debacle is take place in all rivers in England , heavy wind from south to north.,…}

2: Extract the meteorological words (gfw_i)

{Debacle, heavy,…} → { gfw_i }

3: Extract the first letter (w_i).

{D,H,P,C,V,T,E,R,L,J,K,T,H,A,
W,A,I,W,Q,M,Q,Z,M,Z,C,G,P,C,
V,T,V,F,Q,B,A,J,K,T,G,Q,W,A,I,
X,P,M,Q,Z,N} → { w_i }

4: n_i is the ASCII value of (w_i)

{3,7,15,2,21,19,21,4,17,11,9,10,1
9,7,0,22,0,8,22,14,13,16,25,12,4,2
,6,15,2,20,19,20,5,16,10,9,10,19,6
,0,22,0,8,23,15,12,16,25,13} → {
ni}

5: Extract the LSB's for each (n_i).

{1000010110100111001110100000100010011000011111001}

6: Convert each eight ($mbit_i$) into a letters "which represents the secret text information"

{1000010,1101001,1100111,0100000,1000100,1100001,1111001}
→ {Big day}

7: End

6. Summary and Conclusions

All the previous methods for "natural language or text" steganography, either attempts to alter the appearance of text elements, such as lines, words, characters, or attempts to alter the semantic and/or syntactic structure of sentences. The restrictions of natural language steganography method, is a result of its structure. So, methods like, LSB's, DCT, and FFT, is impossible to utilized. For reasons of secure communications, many steganography application methods try to generate the digital cover

*Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq

carrier, and then used it for secret information hiding. If the cover carrier is a txt file which, uses any natural language, a NLG is used for this process. If the letters of the hidden message or the bits of hidden message embedded with random number generator that generate random number each time is used with any NLG, the expected output is a text file written in any natural language. The used NLG system is Forecast Generator (FOG), a weather forecast system that generates bilingual text in English and French. By using any language translator to convert the text language from the English and French to any language such as Arabic language and or use any other NLG system, gives the proposal method its strength.

For this time all steganography methods like, LSB's, DCT, and FFT, can be applied to natural language steganography.

For the used random number generation the input prime number, a, b, should be chosen carefully.

7. References

[1]:Song Y., "Number Theory for Computing", Spring-Flag Berlin Heidelberg New York, 2000.

[2]:Mercan T. , Edward J. and Taskiran E. , "Natural Language Watermarking", Purdue University, West Lafayette, Indiana, 47907.

[3]: Pierre Moulin, and Joseph A, " Information Theoretic Analysis of Information Hiding," O'Sullivan University of Illinois Washington University Beckman Inst., October 1999.

[4]: Neil F., Zoran D., and Sushil J., "Information Hiding Steganography and Watermarking Attacks and Countermeasures", Printed in United State of America, 2000.

*Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq

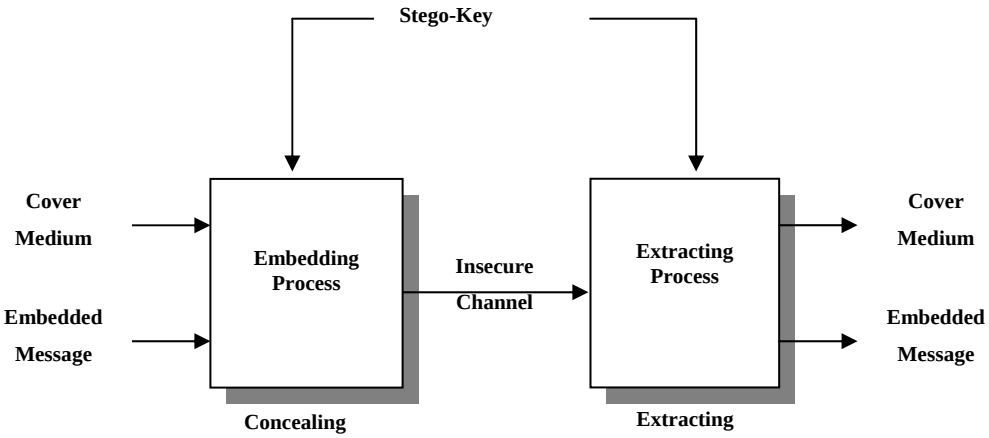

Fig. 1: The General Steganography System

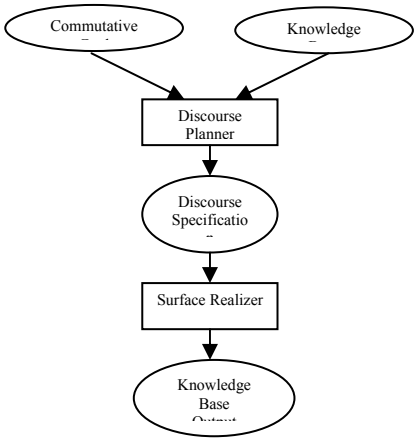

Fig. 2:Components of a typical natural language generation system.

*Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq

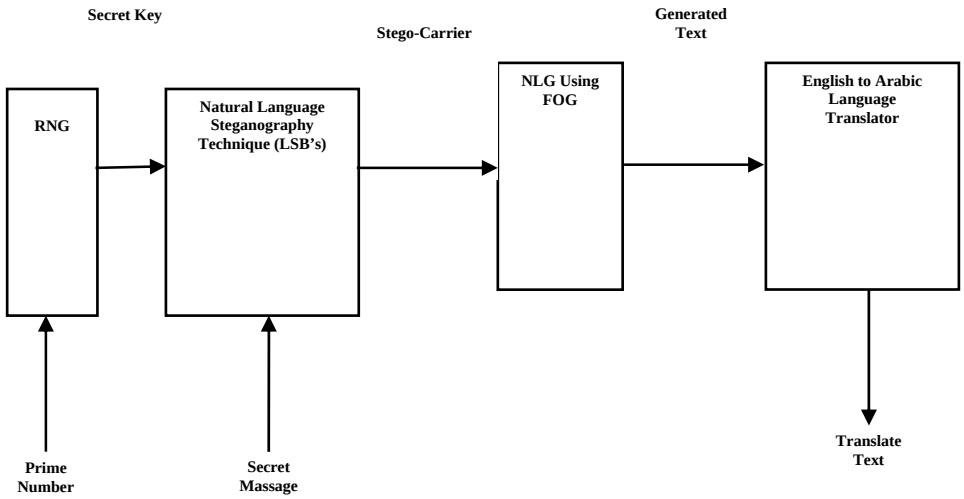

Fig. 3: The proposal technique implementation

*Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq

Extracting Process

*Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq