A Universal Lexical Steganography Technique

Ahmad Alabish, Abdulbaset Goweder, and Anes Enakoa

***Abstract*—A literally meaning of Steganography is "covered writing". There are several methods of steganography, these include: Image steganography, Audio steganography, Video steganography and Linguistic steganography which use the cover to hide information. Each method has its own algorithm to embedding secret information inside the media "cover". Linguistic steganography is basically hiding information in a text in such a way without making the text suspicious, so we have to take into our account possible characteristics of natural languages. In linguistic steganography, digital numbers like (0010100101001) data is to be encoded to innocuous natural language text by using synonym. In this paper, English language will be used as an instance of natural languages as we will be concerned with the set of all natural language texts. this research tries to employ a set of all synonyms as a way to hide secret message inside a natural language text. The main objective of this paper is to develop a general technique of lexical steganography to support different natural languages texts and decrease the bits used for encoding and increase the information. An evaluation of the proposed method has been carried out. The obtained results are encouraging and promising.**

***Index Terms*—Seganography, lexical, linguistic steganography, information hiding, word choice-steganography.**

I. Introduction

With the expand use of computers over the networks and growth of the Communications. This has led to especial security method in computer networks the security for the massage and information has become a necessity for transmitting information. There are two techniques designed to make messages and information transmission more secure through computer networks. These techniques are: cryptography and steganography both techniques are used to hide information.

The meaning of Steganography is "covered writing", steganography embeds information into a file which can not easily be ruined, but no message exactly is indestructible, so it is to take a piece of information and hide it within a cover. The cover might be some computer files like images, text, sound and videos, For example, when the message is hidden inside an image or a sound file in such a way, people can not figure out that there is extra information inside the image or the sound file, While they are looking at the image or listening to the sound.

- Several methods of steganography use the cover to hide information. Each method is requested by an algorithm to embedding secret information inside the media "cover". To protect embedding process, the algorithm sometimes uses keyword so the person that knows the secret keyword can access the secret message on the media. In the next sub-section, steganogrophy methods are presented and discussed.

A. Steganogrophy Methods

- Image steganography: This is the most common method used to hide secret messages because it is simple to implement without changing the properties of the image. So, it is difficult for people to distinguish between the original image and the modified image after embedding a secret message. The images are represented as arrays of numbers. These numbers represent the light intensity of each pixel. There are two types of digital images. Either 8-bit or 24-bit digital images. There are different techniques used for hiding data inside an image. These are: the least significant bit (LSB), masking and filtering, and the algorithm and transformation.
- Audio steganography: Hiding data inside an audio file (frequencies which human can not hear) can be done in the time domain as will as in the spectral domain. There are many audio steganography methods based on embedding capacity and robustness. These include: low bit encoding, Spread spectrum, and Perceptual masking.
- Video steganography: This method is similar to the image steganography method and there is no much difference between these two methods. We can say the video steganography is a derivative of image steganography, because the video is a series of images that are transmitted according to a certain way.
- Linguistic steganography: The linguistic steganography is basically hiding information in a text. In linguistic steganography, machine-readable data is to be encoded into innocuous natural language text. According to this method, we insert a word into an innocuous natural language text as a simple carrying information without making it suspicious. The linguistic steganography method is safer than other methods. This reason has motivated us to carry out a research in this area.

In this paper, we will be concerned with the set of all natural language texts. The proposed technique attempts to employ set of all synonyms as a way to hide a secret message inside a natural language sentence, so that it does not sound suspicious.

II. Lexical Steganography

The lexical Steganography is symbolic. This approach is

Manuscript received August 14, 2012; revised October 1, 2012.

Ahmad Alabish and Anes Enakoa are with the College of computer technology /computer science, Zawia, Libya (e-mail: zawia_ej@yahoo.com, anis_annacoa @yahoo.com).

Abdulbaset Goweder is with the High institute of surman /computer science, Surman, Libya (e-mail: agoweder@yahoo.com).

called a substitution meaning-preserving, if it never changes the whole meaning is traditionally established the relation between the lexical and synonyms (Richard Bergmair 2004)

They refer to a set of words that have the same meaning by a symbol, for example:

C = {Tripoli is a nice little city,
Tripoli is a fine little town,
Tripoli is a great little
Tripoli is a decent little
Tripoli is a wonderful little town}

The above set of sentences can be encoded using known synonyms when there are two distinct sets of synonyms. The first set has five synonyms, while the second set has only two synonyms. The previous set of sentences can be re-written according to the following:

Tripoli is a little
Nice
Fine
Great
Decent
Wonderful
City
Town

All we need to do is to assign binary codeword to each word choice, where we can make word choice in the secret message according to code words.

Tripoli is a little
00 Nice
01 Fine
10 Great
11 Decent
?? Wonderful
0 City

To apply this encoding on the message, the secret message 110 encodes the sentence “Tripoli is a little decent city “. A problem arises using this method that on block codes each word choice is encoded for fixed number of bits, so, we only use a power of 2 for number of word choices in each set of synonym word.

III. Methodology

Using the lexical steganography, we could embed many binary numbers in a natural language text without making it suspicious, but the capacity of information is low and the density of bit is high. So, we try in our paper to increase the capacity of information and safe more bits to present the secret message.

The English alphabet set is represented by a set of letter codes. The English alphabet consists of 26 letters. To represent the English alphabet plus a space character, we need 5-bit letter code. The five bits can represent up to 32 letters which obtained by powering 2 to 5 bits. Table I depicts the English alphabet plus the space character and their binary codes.

TABLE I: Depicts the English Alphabet Plus the Space Character and Their Binary Codes.

Letter	letter code
A , a	00000
B , b	00001
C , b	00010
E , e	00011
D ,d	00100
F , f	00101
G , g	00110
H , h	00111
I , i	01000
J , j	01001
K , k	01010
L , l	01011
M , m	01100
N , n	01101
O , o	01110
P , p	01111
Q , q	10000
R , r	10001
S , s	10010
T , t	10011
U , u	10100
V , v	10101
W , w	10110
X , x	10111
Z , z	11000
␣	11001

Looking at table I and according to the binary codes, the English set of letters can be classified into three different sub-sets or groups. The first sub-set contains the first eight upper letters (*A..H*) where the change occurs in the first 3 bits of the letter code while the last two bits are kept unchanged. In this case, the first sub-set can be represented only by 3-bit letter codes instead of 5-bit letter codes. This leads to save 2 bits for each letter in the first sub-set.

The second sub-set is the second eight middle letters (*I..P*) where the change occurs in the first 4 bits of the letter code, while the last bit is kept unchanged. This group of letters can be represented only by 4-bit letter codes instead of 5-bit letter codes. This means that 1 bit can be saved for each letter in the second group. It is known that the most frequent English letters used to form English words are the English letters (*A..P*) which represent the first two subsets. The third sub-set consists of the last ten lower letters in table 1 (*Q..Z*) plus the space character.

Since the English language is rich in vocabulary which means that many adjectives have several synonyms, this has led to propose a method that is thoroughly based on synonyms usage and gave the proposed method some flexibility.

Table II shows some English adjectives with their synonyms which are used in the proposed method.

TABLE II: SOME ADJECTIVES AND THEIR SYNONYMS

Boring , deadening, dull, ho-hum, irksome, slow, tedious, tiresome, wearisome
Brave courageous fearless desperate, heroic, gallant, lionhearted, stalwart, stouthearted, valiant, valorous
Careful blow-by-blow, cautious, conscientious, painstaking, scrupulous, detailed, elaborate, elaborated, minute, narrow, overcareful, too-careful, studious
Charming , magic, magical, sorcerous, witching, wizard, wizardly, supernatural, influence, tempt
apparent broad, unsubtle, clear-cut, distinct, trenchant, limpid, lucid, luculent, pellucid, , perspicuous
Clever artful, smart, intelligent, adroit, ingenious, cagey, cagy, canny, apt, cunning
Dark black, pitch-black, pitch-dark, aphotic, caliginous, Cimmerian, crepuscular, darkened, darkening, darkling, glooming, gloomy, gloomful, lightless, unilluminated, unlighted, dusky,

Several synonyms of any given word can be used to carry information.

The difficult part of the proposed method is how to represent any given alphabetical letter by all synonyms without causing interference that makes the proposed method embed information inside any general natural language text such as:(news papers, catalogs, advertisement, etc…).

The difficultly takes place as a result of limited and insufficient number of synonyms for each given adjective. It is hard to find an English adjective that has at least 26 synonyms to represent the English alphabet.

To figure out this problem, the set of adjectives synonyms has to be duplicated or tripled in order to cover the remaining letters.

For example, if an English adjective such as clever has about 9 synonyms. This set can not cover all English letters. It only covers the first 10 letters (*A..J*). To cover the whole alphabet, this set of synonyms has to be tripled to cover the letters.

Table III shows different sets of synonyms. A thick horizontal line drawn in table 3 indicates that the end point of the set of synonyms and the start point of copyness. This repetition helps us cover the rest of English letters.

According to this approach, there are two or more different letter codes carried by the same word in a set of synonyms. To identify the letter code that represents a given letter, an illustration example will be presented as follows:
Example:

A natural language text (e.g.: English text) is randomly extracted from any particular essay. A secret message has to be embedded inside the selected text. Suppose that the selected text is” I have a clever friend. His name is Ahmed. He has a car. Its color is black.”

Our task is to embed the secret message “ok” inside the above selected text.

To embed a secret message inside a text, an encoding algorithm is required. This is referred to as an encoder. On the other hand, a decoding algorithm is needed to retrieve the original secret message from the original text. This function is referred to as a decoder.

The encoder: the main objective of the encoder is to search sequentially a word from the selected text and look for it inside a database. If it is found in the database, then replace it with its synonyms according to the letter code obtained from the secret message .

In our case, the letter code which represents the first letter "o" in our secret message is "1110"

TABLE III: SHOWS DIFFERENT SETS OF SYNONYMS

L.code	Synonym1	Synonym2	Synonym3	Synonym4
000	Boring	Brave	Dark	Clever
001	deadening	Courageous	Black	artful
010	Dull	Fearless	pitch-black	smart
011	irksome	Desperate	Caliginous	intelligent
100	slow	Heroic	Cimmerian	adroit
101	tedious	Gallant	Crepuscular	ingenious
110	tiresome	Stalwart	Lightless	cagey
111	wearisome	Valiant	Unlighted	canny
1000	Boring	Valorous	Dusky	apt
1001	deadening	Brave	Darkling	cunning
1010	Dull	courageous	Dark	Clever
1011	irksome	Fearless	Black	artful
1100	slow	Desperate	pitch-black	smart
1101	tedious	Heroic	Caliginous	intelligent
-				
11001				

The encoder: the main objective of the encoder algorithm is to fetch sequentially words from the selected text. For each fetched word, look for it within a particular database. If a searched word is found in the database, then a set or group of synonyms has been identified and this means that synonyms categorized into sets or groups. Once the set of synonyms has been recognized, the next step is to identify the letter code which represent the first letter in the secret message. In our case, the first letter of our message is "o" which is equivalent to the letter code "1110". The encoder will now pick up the adjective which is equivalent to the letter code from the database. If the chosen adjective is identical to the one found in the text, then nothing to be done. Otherwise, the first adjective found in the text will be replaced by the synonym that is equivalent to the letter code of the first letter in the secret message.

In our example, the first adjective "clever" would be replaced by the synonym "adroit". Our text becomes: "I have adroit friend. His name is Ahmed. He has a car. Its color is black"

The encoder will carry on the same process by looking for the second adjective in the text then fetches it and searches for it in the database to identify its group. Once, the group has been recognized, the encoder reads the second letter in the secret message and tries to find its letter code, then identify the adjective or its synonyms that is equivalent to this letter code.

In our case, the second adjective from our text is "black" would be replaced by the synonym "dark" which is equivalent to the letter code "1010".

Our text becomes:"I have adroit friend. His name is Ahmed. He has a car. Its color is dark".

The encoder algorithm will repeat the previous process until no more adjectives in the text are found.

The decoder: it receives the sent text as a modified file which will be opened and read. The decoder reads words sequentially then looks each word in the database. If the searched word is found, then its equivalent letter code is extracted. As mentioned earlier, our database contains adjective repeated several times. Consequently, the decoder would extract several distinct letter codes for the same adjective because of repetition.

A set of letters equivalent to letter codes would be produced by the decoder for each searched adjective.

In our example, the following set of letters will be generated by the decoder for first searched adjective which is "adroit".The first set is "{*e*,*o*,*z*}.

For the second searched adjective which is "dark", the decoder will produce the following set of letters. The second set is: {*a*,*k*,*v*}.

Finally, the decoder algorithm will try out all possibilities (combinations) among the produced sets of letters, then looks up each possibility in a dictionary of words to retrieve the hidden secret message "ok".

In our case all possibilities are tried as follows:

The first set is "{*e*,*o*,*z*}

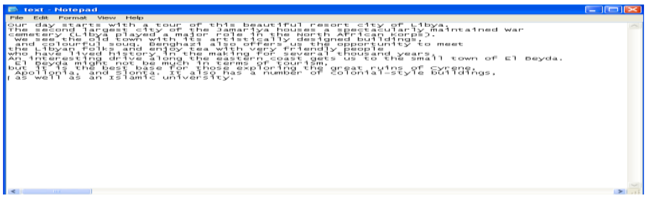
The second set is: {*a*,*k*,*v*}.

This process will generate about 9 possible cases which will be looked up in a dictionary of words. As a result, only one possible case will be found in the dictionary and would be presented by the decoder as the hidden message "ok".

IV. Evaluation

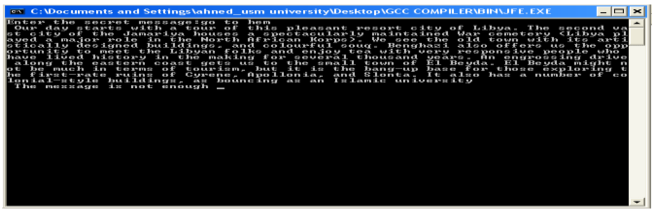
The proposed method has been evaluated for its correctness using different sets of test data. We have chosen randomly several secret massages and attempted to embed them into randomly chosen text. Each secret massage has been embedded into a text correctly by the encoder algorithm without making the original text suspicious. The process of encoding was successfully accomplished for all test secret messages. On the other hand, the decoder algorithm has also been assessed for its correctness. The results have shown that all sent texts are received embedded with secret messages showing no sign that these text are suspicions.

In addition, the decoder algorithm was able to extract hidden secret message correctly from the received texts. Samples of text and secret messages used to evaluate the proposed method are given as shown in Fig. 1, Fig. 2, and Fig. 3



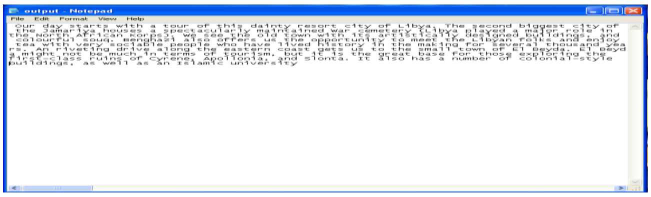
```
Our day starts with a tour of this beautiful resort city of Libya.
The second largest city of the Jamariya houses a spectacularly maintained war
cemetery (Libya played a major role in the North African Korps).
 We see the old town with its artistically designed buildings,
 and colourful souq. Benghazi also offers us the opportunity to meet
the Libyan folks and enjoy tea with very friendly people
who have lived history in the making for several thousand years.
An interesting drive along the eastern coast gets us to the small town of El Beyda.
 El Beyda might not be much in terms of tourism,
but it is the best base for those exploring the great ruins of Cyrene,
 Apollonia, and Slonta. It also has a number of colonial-style buildings,
 as well as an Islamic university.
```

Fig. 1. Original text.


C:\Documents and Settings\ahned_usm university\Desktop\GCC COMPILER\BIN\JFE.EXE

```
Enter the secret message:go to hem
 Our day starts with a tour of this pleasant resort city of Libya. The second va
st city of the Jamariya houses a spectacularly maintained War cemetery (Libya pl
ayed a major role in the North African Korps). We see the old town with its arti
stically designed buildings, and colourful souq. Benghazi also offers us the opp
ortunity to meet the Libyan folks and enjoy tea with very responsive people who
have lived history in the making for several thousand years. An engrossing drive
 along the eastern coast gets us to the small town of El Beyda. El Beyda might n
ot be much in terms of tourism, but it is the bang-up base for those exploring t
he first-rate ruins of Cyrene, Apollonia, and Slonta. It also has a number of co
lonial-style buildings, as bouncing as an Islamic university
 The message is not enough _
```

Fig. 2. Secret message:"Go to him".



```
 Our day starts with a tour of this dainty resort city of Libya. The second biggest city of
 the Jamariya houses a spectacularly maintained War cemetery (Libya played a major role in
the North African Korps). We see the old town with its artistically designed buildings, and
 colourful souq. Benghazi also offers us the opportunity to meet the Libyan folks and enjoy
 tea with very sociable people who have lived history in the making for several thousand yea
rs. An riveting drive along the eastern coast gets us to the small town of El Beyda. El Beyd
a might not be much in terms of tourism, but it is the great base for those exploring the
first-class ruins of Cyrene, Apollonia, and Slonta. It also has a number of colonial-style
buildings, as well as an Islamic university
```

Fig. 3. Modification file

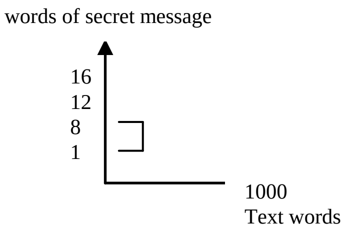


Fig. 4. Secret messages of lengths between 1 to 8 words.

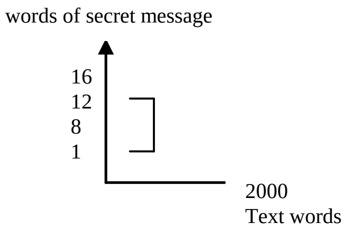


Fig. 5. Secret messages of lengths between 1 to 12 words

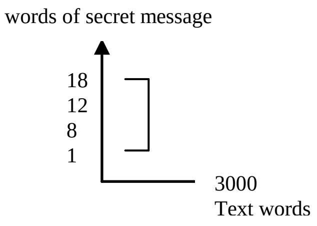


Fig. 6. Secret message of lengths between 1 to 18 words.

Evolution of the proposed method in terms of efficiency has shown that large enough text containing sufficient distinct adjective is required to be able to embed a secret message with a specific length. As the secret massage gets larger, the text used to embed this message has to be large enough. This leads to slowness of the proposed method as secret massages become lengthy. Beside, in terms of resources, more space is needed as secret messages get larger.

Fig. 4, Fig. 5, and Fig. 6 show length of the texts in terms of words needed to cover different lengths of secret messages. Fig. 4 shows secret messages of lengths between 1 to 8 words need about 1000 words of text to embed these secret massages. Fig. 5 shows that 2000 words are needed for secret messages of lengths between 1 to 12. Fig. 6 shows that 3000 words are needed for secret message of lengths between1 to 18.

V. Conclusion

Linguistic steganography has become one of the important methods in hiding information on the cover, because it uses natural language text to hide information. Consequently, this type of cover is harder to attack than other covers. In this paper, a method for embedding a secret message into any text containing several distinct adjectives has been proposed. A secret message can be embedded into any text without changing the features of the text being used. The proposed method is general and can be used for all natural languages. It is independent of a human language the proposed method is based on applying synonyms of an adjective to cover a secret message. Each set of synonyms has to cover the alphabet. When the text being used to cover a secret massages contains sufficient number of adjectives, then long secret massages would be embedded easily. The results of our experiments have shown that the proposed method was successfully able to embed secret message with different lengths into different texts. Besides the method was successful in extracting and retrieving the hidden secret massage out of the text.

Finally, it can be concluded that the results we have obtained are encouraging and given us high motivation to carry out a research in this area.

VI. Future Work

An improvement in terms of efficiency of the proposed method has to be carried out.

References

[1] R. Krenn. *Steganography and Steganalysis*.

[2] R. Bergmair, "Towards linguistic steganography: A systematic investigation of approaches," *Systems and Issues*, 2004.

[3] K. Winstein, *the Word Choice Hash*.

[4] N. F. Johnson, *Introduction to Steganography: Hide Information (, PH.D), Center for Secure Information System*.

[5] K. Winstein. Tyrannosaurus Lex|An Implementation of Lexical Steganography. [Online]. Available: http://www.imsa.edu/~keithw/tlex

[6] K. Bennett, *Linguistic Steganography: Survey, Analysis, and Robustness Concerns for Hide Information in Text*.

[7] J. J. Chae and B. S. Manjunath, *Data Hiding in Video.*

[8] K. Sayood, *Introduction to Data Compression* (third edition).