IMPROVING N-GRAM LINGUISTIC STEGANOGRAPHY BASED ON TEMPLATES

Alfonso Muñoz, Justo Carracedo Gallardo
Universidad Politécnica de Madrid, Escuela de Ingeniería Técnica de Telecomunicación
Departamento de Ingeniería y Arquitecturas Telemáticas, Carretera de Valencia Km. 7, 28031 Madrid, Spain

Irina Argüelles Álvarez
Universidad Politécnica de Madrid, Escuela de Ingeniería Técnica de Telecomunicación
Departamento de Lingüística Aplicada a la Ciencia y a la Tecnología
Carretera de Valencia Km. 7, 28031 Madrid, Spain

Keywords: Linguistic Steganography, Automatic generation of stegotext, Stelin.

Abstract: The automatic generation of stegotexts is an area of research within linguistic steganography with a great interest. Different proposals have been published about the development of stegotexts with statistical and linguistic validity in natural language but, unfortunately, the stegotexts generated following some of these procedures suffer from different types of lexical, syntactic, semantic problems or problems of global coherence that would eventually lead the automatically generated texts to be detected by a human reader. This article presents an improved implementation of an algorithm of N-Gram statistic imitation and introduces a new concept of manual edition of the stegotexts generated, based on the idea of templates. This procedure permits the creation of high quality stegotexts to hide a small quantity of information, hundreds of bits, useful to distribute a symmetrical key, a short message, an url, etc. The procedure developed in the article can be applied to different languages, but here, its usefulness is demonstrated for Spanish. The Stelin tool is presented free at http://stelin.sourceforge.net.

1 IMPROVING N-GRAM LINGUISTIC STEGANOGRAPHY BASED ON TEMPLATES

Using a N-Gram model (the statistics that guides the relation between words and collocations in a language) to conceal information can be very productive today. In the following subsections, a series of improvements applied to Wayner’s model, *mimic function*, (Bergmair 2007) are commented and the concept of manual annotation is introduced. As an example, fragments of stegotexts in Spanish are included. Spanish language was chosen because it is a language spoken by more than 400 million people in the Internet and because it is the native language of the authors. A similar reasoning can be followed for other languages as for example, English.

1.1 Improving Lexical and Grammatical Quality of the Generated Stegotext

The tests carried out indicate that it is more useful to measure the statistics of every "word" of the training text instead of every "letter" (in Wayner's original idea). This change implies a lexical and grammatical improvement of the generated stegotexts quality, at least in Spanish language. In the same way, the use of the punctuation marks (.:; ¿?!, etc.) as individual words, separating them if they are joined to another word, not only influences the grammatical improvement of the generated stegotext but it also facilitates the concealment of more information. For example, if we consider the phrase: "The sky is cloudy.", it is more probable that possible following words are written after the full stop rather than after the word "cloudy.", and following Wayner's idea, if the number of following words is higher, more

Muñoz A., Carracedo Gallardo J. and Argüelles Álvarez I. (2010).
IMPROVING N-GRAM LINGUISTIC STEGANOGRAPHY BASED ON TEMPLATES.
In *Proceedings of the International Conference on Security and Cryptography*, pages 209-212
DOI: 10.5220/0002918202090212

information could be hidden. Having in mind these principles the following algorithm is defined:

1. Create a root table with all the different existing words in the training text. Every entry of the table will connect to another table (level) that contains the words that can follow it and this way recursively up to an order N (Figure1). With this information make a Huffman tree for each level.
2. A "word" is selected at random from the root table. This selection allows for the same training text to generate different stegotexts.
3. If the word does not have successors, it does not point to another node, another term in the root table is chosen. If the node only has a successor word, this word is added to the stegotext and the following available node is chosen. Information cannot be concealed in this case. If the successor node has several possible words among which to choose, choose one with a binary code (path of the Huffman's tree) that coincides with the information to be concealed. After, the following available node is chosen.
4. If the last node is reached (order n=8, for example, 8 consecutive words) the last word is selected for the stegotext and back to step b). This process is repeated until the stegotext that conceals the information is generated.
5. The receiver needs to build the table of frequencies of the selected training text to know the bits that conceal every word of the received stegotext.

Although this variant generates stegotexts of larger size, it is easier to obtain texts with lexical and syntactic validity (Figure 2). The tests carried out indicate, at least in Spanish language, that in general, an order 7 or higher provides lexical and syntactic reasonable results. Also, following the results of tests carried out, a higher order than 7 or 8 (depending on the training text) can be sufficient. In fact, over this order the generated stegotext does not improve substantially its linguistic appearance whereas its size is substantially bigger.

The importance of the statistical imitation is that on having imitated statistics, the linguistic validity of the training text would be imitated. Depending on the channel used to transmit it, the stegotext should keep the statistical conditions and those linguistic properties typical of the channel not to raise suspicions.

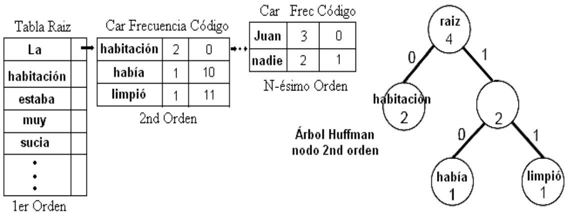

Figure 1: Example of implemented algorithm in Stelin. Mode=word.

1.2 Manual Annotation and Templates

The Stelin tool (Muñoz 2009) developed generates automatically stegotexts in natural language based on the statistical imitation (N-Gram) of one or more training texts, known only by the transmitter and the receiver. The transmitter and the receiver only need to keep the training sources private and a cryptographic key if they want to cipher the information to be concealed; the rest of the system is public.

Imitating the statistics of the words in the training texts leads to stegotexts with better lexical and grammatical quality. Nevertheless, depending on the training source, the generated stegotexts still might have less important grammatical mistakes as for example question marks that are opened but are not closed, as well as possible problems of global coherence which is a main problem in the science of linguistic steganography. An interesting idea, while an algorithm that solves the above mentioned problems is obtained, would be editing the stegotexts generated after their generation to correct those mistakes. In general, this procedure has the fundamental problem of synchronization with the receiver, that is to say, introducing changes in the generated stegotext implies that the receiver must know what changes have been made in order not to have problems in the recovery of the secret information. Up to what is known, solutions have not been published in this respect, though there are solutions that permit to improve the quality of the stegotext in the process of generation (Bergmair 2007). The proposed idea consists of improving the linguistic appearance of the generated stegotext a posteriori without the need that the receiver has any additional information about the changes. This idea can be put into practice with the Stelin tool with a series of considerations and limitations.

Let's think for a moment about the structure in the example in Figure 1. A piece of information is hidden by means of the selection of a word among the available ones in the following level. That is to say, if the word "La" (the) is selected, the following word might be “habitación” (room) (bit 0), "había" (had) (bit 10), “limpió” (cleaned) (bit 11) depending on the bits to be concealed. The receiver would construct the frequency table as the transmitter and would relate the words in the stegotext received with the generated table. The receiver would know that the words "La habitación", “La había” or “La limpió” hidden one bit (0) or two (10-11) respectively.

The receiver might operate not considering the words that are read until one finds one of the possible words at the level of the linked list where we are. This small not published detail, permits the a posteriori manual improvement of generated stegotexts by the transmitter without involving the receiver. The only condition is that the transmitter can use any word that one cannot find in the following level so that the receiver does not lose synchrony. That is to say, in the previous example among the words that conform the pairs "La habitación”, “La había”, “La limpió”, one might use any word that was not “habitación”, “había” or “limpió”. This way, the transmitter can correct possible grammatical mistakes and improve the coherence of the text without any need that the receiver has this information. For example, if we conceal a small information of 126 bits, using an order 9 and Antonio Machado's complete poetry as training text we would obtain among the possible stegotexts one like the following one:

*[…] suena el rebato de la tarde en la arboleda**!** **M**ientras el corazón pesado. El agua en sombra pasaba tan melancólicamente, bajo los arcos del puente al ímpetu del río sus pétreos tajamares; […]*

Figure 2: Stegotext that conceals 126 bits. Order 9. Clave: Alfonso. Training source Antonio Machado's complete poetry.

As it can be observed, this stegotext has a series of small grammatical mistakes (we leave aside problems of global coherence), for example in the fragment*“la tarde en la arboleda! Mientras”* (the evening in the grove! While[…]), a word with a capital letter that is not preceded by a full stop and a question mark closed but not opened before. To solve these problems the Stelin tool generates a template with the possible words in every level, so that the transmitter could select the words to be added among the words of the stegotext, words that will be rejected by the receiver. For example, we select the fragment: “*de la tarde en la arboleda! Mientras el corazón*” (of the evening in the grove! While the heart"). Let's see an example of the generated template:

[WORD:en][muerta][flota][.][bella][roja][en][,]
[sobre][arrebolada][y]
[WORD:la][sus][la]
[WORD:arboleda] [arboleda]
[WORD:!] [!]
[WORD:Mientras] [Mientras]

Bearing this in mind we edit the fragment. Among the multiple possible options we choose the following one:

*“tarde en la **dulce** arboleda**, ¡qué sensación**!**.** Mientras el corazón” (*late in the sweet grove, what a sensation!. While the heart)

This so simple way the quality of the generated stegotext can be improved substantially, for example, solving problems derived from punctuation marks or others. The modification of the stegotexts by this procedure needs a bit of practice. Thankfully in the texts in natural language in Spanish (and in other languages) the words have certain positions, that is to say, it is more probable that certain words appear after others (Zipf law) and it is more probable that more words appear after some others. In general, there would be few words after which it would be difficult to choose a new word because there are many options in the level of the corresponding table, and many words after which few options will exist so there will be easier to add new words. For example, in texts in Spanish language words as ***de, la que, el, en, y, a, los,*** etc, are more probable, then it is more probable that there are more words that can appear with these. If we concentrate on the previous example “*de la tarde en la arboleda! Mientras el corazón*” (of the evening in the grove! While the heart ") it would turn out to be trivial to add information before the article “el” (the) but more difficult to find words after this word and before the word “corazón” (heart). Bearing this principles in mind can help to reduce the time and the positions where it is better to work to correct the possible grammatical mistakes. So it is possible to automatically generate stegotexts, at least in Spanish language, and to correct the mistakes manually without involving the receiver and obtaining stegotexts with a quality more than average.

Nowadays the concealment of information with the Stelin tool has a capacity of approximately 1 WORD-2 BIT (it depends on the training text and on the information to be concealed), this capacity of concealment diminishes depending on the number of words that are added by means of manual edition. To conceal information over a hundred bits the generated stegotexts will be of large size (it depends on the size and on the "quality" of the training texts) and therefore the manual edition will take a considerable amount of time. Nevertheless, another function of the tool is possible taking advantage of the manual edition characteristics. The procedure would consist in the generation of "bad quality" stegotexts but with the quality of occupying little space. Provided that we can modify the resultant stegotext we might use a very low order (what would produce a bad quality stegotext) to generate a small size stegotext to improve it manually later. In addition, it would turn out to be simpler, when the generated stegotext had fewer words, to make a substantial improvement of the global coherence (Figure 3 and 4).

está bien visto bastante por la puerta del salón del trono. El hombre no se podía dejarlas salir de la mano, y la cabeza, la idea de si no me pareció tan grande como si no me apetecía estar ni nada de los que le hizo un gesto de asombro, Marla se

Figure 3: Created stegotext that conceals 126 bits. Order 3. 52 words. 2,4 bits/Word. Training source: Antonio Machado's Complete Poetry.

está bien visto *sus modales por* ***bastante*** *gente, especialmente cuando luce sus encantos* ***por la puerta del salón del trono. El hombre no se*** *esconde, es más* ***podía*** *ser el amante perfecto enamorando a las mujeres sin* ***dejarlas*** *decidir. Es famosa la doncella que no quiso salir* ***de la mano*** *con él****, y*** *orgullosa levantó* ***la cabeza, la*** *valiente Marla. Una* ***idea*** *disparatada* ***de*** *asombrosa valentía. Soy el narrador y* ***si no*** *lo escribo morirá está hazaña, una hazaña que* ***me pareció tan grande como si*** *un animal hablara. Es extraño,* ***no me apetecía estar*** *callado* ***ni*** *guardar silencio, [...]*

Figure 4: Stegotext of the Figure 3 with manual annotation and coherence. 128 words. 1 bit/word.

her modals are well seen by most people, especially when she shows her charms through the throne-room door. The man does not hide, what is more, he could be the perfect lover, making women fall in love without letting them decide. Famous is the lady who did not want to walk hand in hand with him and proud raised her head, brave Marla. An absurd idea of amazing courage. I am the narrator and if I do not write it, the heroic deed will die, an heroic deed that I thought to be so incredible as if an animal talked. It is strange, I did not feel like being quiet or keep silent, [...]

Figure 5: Stegotext of the Figure 4 in English. Approximate Translation.

As it can be observed, even generating bad quality stegotexts (low order or as a result of a bad selection of the training source) the use of templates can help not only to correct grammatical mistakes but also to provide coherence to the text. As it can be seen, if the chosen order is high (7 or more) the correction of the texts will be faster (as fewer mistakes exist) though the generated stegotext will be larger. On the contrary for low orders (for example, N=3) the text will be smaller but the correction will need of more time and possibly it will be necessary to add more words.

2 CONCLUSIONS

Linguistic steganography is a science that in the last years is arousing the interest of the scientific community due to its enormous potential.

This article has presented the Stelin tool that implements a series of improvements of a stegotext generation automatic algorithm based on N-GRAM statistical imitation. A new concept has been introduced, that of stegotexts manual correction that allows the correction of grammatical mistakes and even to improve the coherence of the generated stegotexts by means of the use of templates.

In the practice good quality stegotexts are obtained but it is necessary to invest a lot of time (it depends on the “writing skills” of every transmitter) to conceal more than one hundred bits (cryptographical keys, urls, mobilization messages, GPS coordinates, etc) by means of manual edition (more time for low order). In any case, the quality of the stegotext can always be modified and perfected by the transmitter. If the information to conceal is important might be that the invested time was worth.

The transmitter and the receiver must share a cryptographic key if they want to cipher the information and one or more training texts. In the case that the training texts were compromised, they would always have the option to change without any problem choosing new texts with which to work.

Today it is not clear if it is possible to obtain secure automated procedures of linguistic steganography that allow to conceal thousands of bits. Meanwhile Stelin can be a good tool to obtain stegotexts of a reasonable quality.

REFERENCES

Bergmair, R., 2007. A comprehensive Bibliography of Linguistic Steganography. S*PIE International Conference on Security, Steganography, and Watermarking of Multimedia Contents*, volume 6505, *January 2007.*

Muñoz, A., 2009. The Stelin Tool http:// stelin.sourceforge. net