Researcher's Perspective of Substitution Method on Text Steganography

Fawwaz Zamir Mansor, Aida Mustapha, Noor Azah Samsudin

Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia (UTHM), 86400, Parit Raja, Batu Pahat, Johor, Malaysia

Corresponding author: aidam@uthm.edu.my

Abstract. The linguistic steganography studies are still in the stage of development and empowerment practices. This paper will present several text steganography on substitution methods based on the researcher's perspective, all scholar paper will analyse and compared. The objective of this paper is to give basic information in the substitution method of text domain steganography that has been applied by previous researchers. The typical ways of this method also will be identified in this paper to reveal the most effective method in text domain steganography. Finally, the advantage of the characteristic and drawback on these techniques in generally also presented in this paper.

1. Introduction

Steganography can be defined as associated knowledge of hiding the messages via medium of data to become invisible and undetectable for human sense. Secure privacy information is critical point of steganography in applying performance as a part of information hiding. The implementation of steganography itself, the methods can be divided into two categories; steganography in medium of image, audio, video and other digitally invisible code named technical steganography. Therefore, this paper is specifically focusing in linguistic steganography using substitution method.

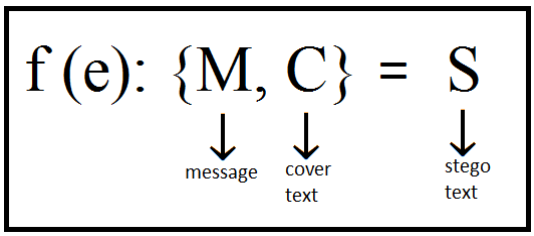

Figure 1: General equation of steganography in text domain

Generally, the function of steganography in text domain basically can be formulized using equation showed on Figure 1. Basically, this process analogically can be illustrated using Prisoner's Problem. The analogy is represented in Figure 2, Alice is sending an original text (M) along with a cover message (C) in order to process embedding known as stego text (S) containing a stego key (K). Firstly, apply the invertible function f (e): {M, C} → S. Alice can plan an original text (M) using a stego key (K) through e (M, C) = S. Hence, S as stego object and it is invertible function, Wendy will not

discover the message as suspicious thing. Then, Bob will figure out e − l(s) = {M, C} in order to retrieve original text M and cover message C with a stego key K for decoding the process use function. The process of embedding and extracting the information of stego text S will be known by Wendy, so that, we can clarify the process steganography is successful.

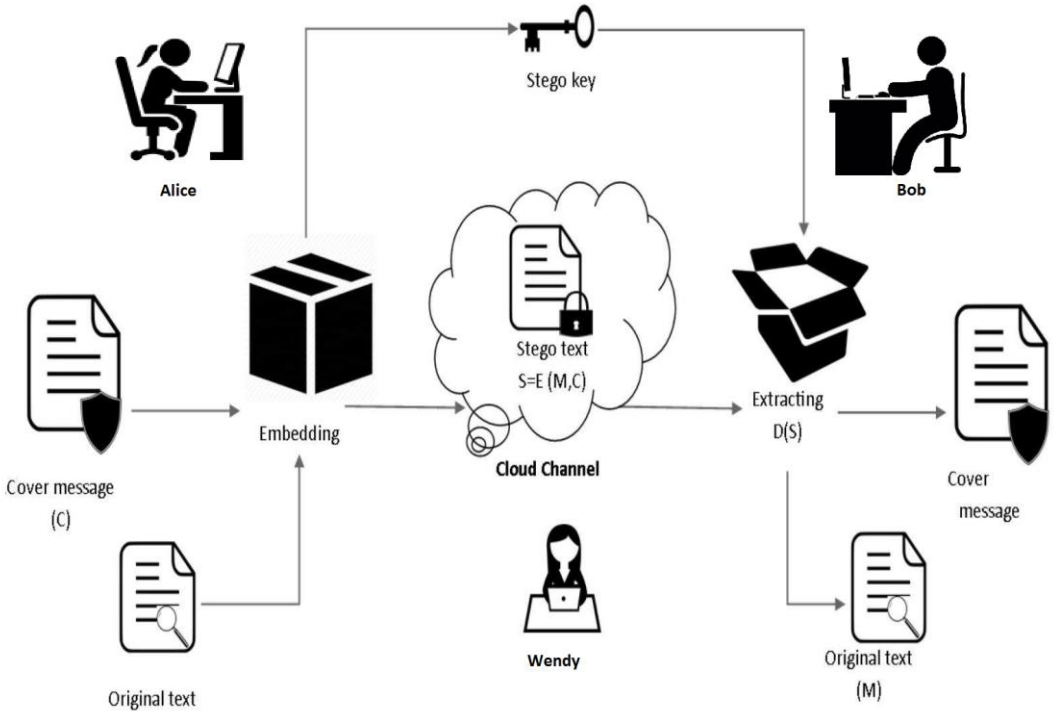

Figure 2: A typical steganography processes

Based on that idea, the general objective is signifying the development of steganography technique about technique steganography in medium of text that obtain from past researchers' effort. The implementation of steganography in text domain can be classified into two categories; format based steganography and linguistic steganography. Therefore, the main concentration of this study is about linguistic steganography and also discussing about text steganography for comparison, with objective to signify the classification of technique. For linguistic steganography, it covers messages which modified the information that encoded the massage based on order of linguistically [1]. Whereas, format based steganography is used in covering messages which manipulated the component in text such as, feature in word, space, line and any other character in sentence of text.

Linguistic steganography is a set of methods and techniques that permit the hiding of any digital information within texts based on some linguistic knowledge [2] [3] that will provide more detailed definition. It not only requiring the stenographic cover be composed of natural language text or some sort, but the text itself is either generated to have a cohesive linguistic structure, or that the cover text is natural language text to begin with.

Furthermore, natural language steganography or linguistic steganography is a technique which aims to hide secret information in text document by manipulating the semantic and/or syntactic structures of the sentence [4]. From the definitions above, we can conclude that the main idea of linguistic steganography is to hide secret message using linguistic knowledge.

2. Previous Study of Substitution Method on Text Steganography

Linguistic steganography can cover the hidden message concerning language of word and order modification linguistically. The implementation of this technique typically uses substitution method to enhance the linguistic steganography. Table 1 below show the previous studies of researcher that have been applied.

Table 1: List of substitution methods in text steganography

	Scholar Author	Method	Substitution	Linguistic (manipulation of language)	Format (manipulation of structure)
1	(Topkara et al, 2006)	Quantifiably resilient watermarking of natural language text through synonym substitutions	√	√	
2	(Chand et al, 2006)	Exploiting linguistic features in lexical steganography	√	√	
3	(Shahreza, 2006)	HTML page is the ID attribute	√	√	
4	(Shirali et al, 2007)	Short Message services (SMS)	√	√	
5	(Yuling,2007)	Linguistic Steganography over Chinese text	√	√	
6	(Shahreza, 2008)	Text Steganography by Changing Words Spelling	√	√	
7	(Shahreza, 2008)	Persian/Arabic Unicode Text Steganography	√		√
8	(Muhammad, 2009)	Linguistic Steganography over Malay text	√	√	
9	(Wang et al, 2009)	Emoticon-based Text Steganography in Chat	√		√
10	(Alla & Prasad, 2009)	An Evolution of Hindi Text Steganography	√		√
11	(Shahreza, 2008)	Synonym Text Steganography over British and American English	√	√	
12	(Shahreza, 2010)	Arabic/Persian Text Steganography Utilizing Similar Letters with Different Codes	√	√	
13	(Alameti et al, 2010)	A new approach to Telugu text Steganography by shifting inherent vowel	√		√
14	(Bensaad, 2011)	High Capacity Diacritics-based Method for Information Hiding in Arabic Text	√		√
15	(Prasad et al, 2011)	A New Approach to Telugu Text Steganography	√		√
16	(Roslan et al, 2011)	Sharp-Edges Method in Arabic Text Steganography	√		√
17	(Memon et al, 2011)	A Novel Text Steganography Technique to Arabic Language Using Reverse Fatha	√	√	
18	(Gardiner, 2012)	Text Steganography on Online Chat	√	√	
19	(Wang et al, 2013)	Novel text steganography by context-based equivalent substitution	√	√	
20	(Qi et al, 2013)	Secure text steganography based on synonym substitution	√	√	
21	(Munoz et al, 2013)	Measuring the security of linguistic steganography in Spanish based on synonymous paraphrasing with WSD	√	√	

Based on Table 1, most researchers had studies many kinds of substitution method, but there are some substitution methods were not categorized as linguistic because of the substitution method had change the format of the language itself. So, it proves that, most of substitution methods are linguistic steganography.

There are some advantages in linguistic steganography. Firstly, unlike format based steganography, linguistic steganography especially in synonym substitution technique has own implementation performance; English text using LUNABEL function [5] and high invisibility in Malay linguistic technique [6], simple variant in Chinese text synonym [7] minimalized creating syntax error in English text or Arabic text using context-based [8] [9] [10] and in English text using mark-insertion can achieve maximum cumulative distortion [11]. Secondly, one of the techniques which is traditional synonym substitution enable to hide the hidden message in large capacity [12] [13]. Thirdly, linguistic steganography also has advantage in specific condition linguistic steganography which can be useful in printing text using synonym substitution [14]. Another technique that has this advantage is synonym

replacement which is very useful in Spanish language [15]. Moreover, there are also using HTML as the ID attribute and Short Message Service (SMS) in their technique [16] [17]

Then, they also identified the disadvantages in this technique. Firstly, low security can also be the issue in linguistic steganography [12] and synonym paraphrasing technique in Spanish language is easy to be attack by intruders [15]. Secondly, the limitation implementing of linguistic steganography is only capable in own language because this technique is based on linguistic. This technique has complex algorithm in English text using LUNABEL function [5] and using traditional synonym substitution [13]. In English text, using context-based has incomplete vocabulary [8]. Then, semantic transformation technique has the possibility generates a lot of semantic spam. Plus, in Malay linguistic, the issue is, it was consuming much in process embedding/extracting hidden message [6]. Most researcher also use synonym substitution as their method to hide secret text, but not secret message. The secret message still easily predictable.

For the format based, they use manipulation of substitution text to change the structure on secret message. As example, some of the use object like Persian/Arabic Unicode [18] [19] [20], Hindi text [21], Telugu text [22] [23], or emoticon symbol [24] on the text steganography to hide secret message.

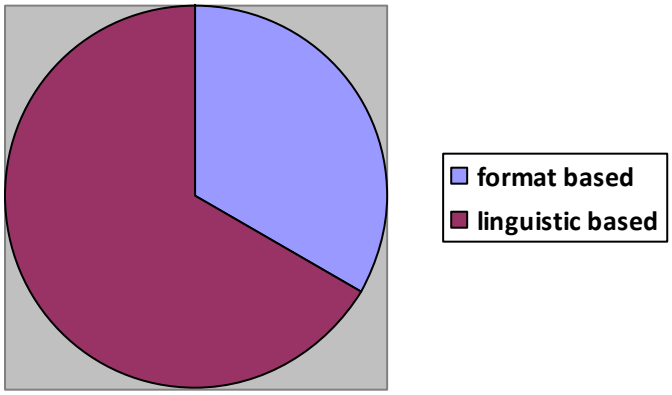

Figure 3: Previous research on format based and linguistic based on substitution method steganography

There is a lot of efforts which have been proposed by previous researcher. The number of the research on linguistic based is much higher than format based refer on Figure 1. Averagely, most widely researched by previous researcher especially in substitution method using synonym substitution.

3. Conclusion

The field studies in linguistic steganography were very interesting to researchers. This paper presented to explore several techniques in steganography methods of linguistic steganography in order to observe the development of these methods previously. In linguistic steganography, the techniques covered the hidden message concerning language of word and order modification linguistically. Moreover, format based and linguistic steganography almost has similar advantage in development, performance, and implement for hiding hidden message. Some techniques have recommended secure protection, high performance, or can embed large amount the hidden message. However, both techniques also have certain issues which seems to be the limitation of those techniques; low security, complex algorithm performance, or time consuming. As the majority of past researcher apply synonym on their studies, the synonym seem like very weak to protect the secret message in stego text, human can predict the secret message easily. In the future studies, researchers could apply

antonym as replacement to the synonym substitution since it can hide real secret message compared to the synonym. At the future, there will be more investigation about on this field will be consider.

Acknowledgement
The author is grateful to Malaysian Ministry of Higher Education (MoHE) under Fundamental Research Grant Scheme (FRGS) to the Office of Research, Innovation, and Commercialization (ORICC) Universiti Tun Hussein Onn Malaysia (UTHM) under Vot 1557.

References
[1] Agarwal, M. (2013). Text steganographic approaches: A comparison. International Journal of Network Security & Its Applications, 5(1):91–106.
[2] M. Chapman, G. I. Davida, (2001) "A Practical and Effective Approchto Large ScaleAutomated Linguistic Steganography", Swiss Federal Institute of Technology Computer Engineering and Networks Laborator :G.I. Davida and Y.Frankel (Eds.): ISC, LNCS 2200, pp 156-165, Springer-Verlag Berlin Heidelberg.
[3] Krista Bennert. (2004) Linguistic Steganography: Survey, Analysis, And Robustness Concerns for Hiding Information in Text. Technical Report TR 2004-13, Purdue CERIAS, May.
[4] H.M. Salman. (2008) " A Natural Language Steganography Technique for Text Hiding Using LSB's" Dept. of Computer Science & Information System Univ. of Tech. Baghdad, Iraq Eng.&Tech. Vol.26, No 3.
[5] Chand, V. and Orgun, C. O. (2006). Exploiting linguistic features in lexical steganography: Design and proof-of-concept implementation. In Proceedings of the Annual Hawaii International Conference on System Sciences, volume 6, pages 1–10.
[6] Muhammad, H. Z., Rahman, S. M. S. A., and Shakil, A. (2009). Synonym based malay linguistic text steganography. In 2009 Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA 2009), pages 423– 427.
[7] Yuling, L., Xingming, S., Can, G., and Hong, W. (2007). An efficient linguistic steganography for chinese text. In Proceedings of Multimedia and Expo, IEEE International Conference, pages 2094–2097.
[8] Shirali-Shahreza, M. (2008, February). Text steganography by changing words spelling. In Advanced Communication Technology, 2008. ICACT 2008. 10th International Conference on (Vol. 3, pp. 1912-1913). IEEE.
[9] Wang, F., Huang, L.and Chen, Z., Yang, W., and Miao, H. (2013). A novel text steganography by context-based equivalent substitution. In Signal Processing, Communication and Computing (ICSPCC), 2013 IEEE International Conference, pages 1–6.
[10] Shirali-Shahreza, M. H., & Shirali-Shahreza, M. (2010). Arabic/Persian text steganography utilizing similar letters with different codes. The Arabian Journal For Science And Engineering, 35(1b).
[11 Topkara, U., Topkara, M., and Attalah, M. (2006). The hiding virtues of ambiguity: Quantifiably resilient watermarking of natural language text through synonym substitutions. In Proceedings of the 8th workshop on Multimedia and security, pages 163–174.
[12] Qi, C., Xingming, S., and Lingyun, X. (2013). A secure text steganography based on synonym substitution. In Conference Anthology, IEEE, pages 1–3.
[13] Gardiner, J. (2012). StegChat: A Synonym-Substitution Based Algorithm for Text Steganography.
[14] Shirali-Shahreza, M. H., & Shirali-Shahreza, M. (2008, August). A new synonym text steganography. In Intelligent Information Hiding and Multimedia Signal Processing, 2008. IIHMSP'08 International Conference on (pp. 1524-1526). IEEE.
[15] Munoz, A., Carracedo, J., and Alvarez, I. A. (2010). Measuring the security of linguistic steganography in spanish based on synonymous paraphrasing with WSD. In Proceedings - 10th IEEE

International Conference on Computer and Information Technology, CIT-2010, 7th IEEE International Conference on Embedded Software and Systems, ICESS-2010, ScalCom-2010, pages 965–970.
[16] Shahreza, M. (2006). A new method for steganography in HTML files. Advances in Computer, Information, and Systems Sciences, and Engineering, 247-252.
[17] Shirali-Shahreza, M., & Shirali-Shahreza, M. H. (2007, November). Text steganography in SMS. In Convergence Information Technology, 2007. International Conference on (pp. 2260-2265). IEEE.
[18] Bensaad, M. L., & Yagoubi, M. B. (2011, April). High capacity diacritics-based method for information hiding in Arabic text. In Innovations in Information Technology (IIT), 2011 International Conference on (pp. 433-436). IEEE.
Chicago
[19] Roslan, N. A., Mahmod, R., & Udzir, N. I. (2011). Sharp-Edges Method In Arabic Text Steganography.
[20] Memon, M. S., & Asadullah, S. (2011). A novel text steganography technique to Arabic language using reverse Fatha. Pak. j. eng. technol. sci, 1, 106-113.
[21] Alla, K., & Prasad, R. S. R. (2009, April). An evolution of Hindi text steganography. In Information Technology: New Generations, 2009. ITNG'09. Sixth International Conference on (pp. 1577-1578). IEEE.
[22] Alameti, S., Pothalaiah, S., & Babu, K. A. (2010). A new approach to telugu text steganography by shifting inherent vowel signs. Networking and Communication Engineering, 2(11), 444-448.
[23] Prasad, R. S. R., & Alla, K. (2011, September). A new approach to Telugu text steganography. In Wireless Technology and Applications (ISWTA), 2011 IEEE Symposium on (pp. 60-65). IEEE.
[24] Wang, Z. H., Kieu, T. D., Chang, C. C., & Li, M. C. (2009, November). Emoticon-based text steganography in chat. In Computational Intelligence and Industrial Applications, 2009. PACIIA 2009. Asia-Pacific Conference on (Vol. 2, pp. 457-460). IEEE.