IMPLEMENTING MALAY LANGUAGE IN LINGUISTIC STEGANOGRAPHY

Amirulikhsan Zolkafli and Roshidi Din
Faculty of Information Technology
Universiti Utara Malaysia, 06010 Sintok, Kedah

Abstract: Steganography, a part of information hiding is an art of hiding information into another medium of information (text, audio, video, image, etc.) with the main purpose is securing information during communication. With the purpose of opening an opportunity to create a communication channel that is covert and subliminal between two parties where by the existence of the particular message being sent and received are kept unknown or innocuous to any possible attacker. Steganography can be divided into two that is technical steganography and linguistic steganography. English have been widely used in linguistic steganography. Therefore, the main objective of this paper is to see whether other languages are compatible to be implement as linguistic steganography. This paper will discuss briefly on linguistic steganography and its system model. The comparison study is done based on linguistic aspects (syntactic and semantic) of several foreign language that have been developed into machine translation system. Chinese, Japanese, and Malay will be compared, in which the characteristic of Malay language will be highlighted in this paper. We will be discussing the purpose of using its language structure to hide secret information in it. This is to show several characteristic of Malay language that could be used in linguistic steganography. Finally, this paper by mentioning the general potential of information hiding and the impact in the form of applications that it could give to Information and Communication Technology (ICT) environment.

Keywords: Steganography, Linguistic, Semantic, Syntactic, Natural language

INTRODUCTION

Steganography, a part of information hiding is an art of hiding information into another medium of information (text, audio, video, image, etc.) with the main purpose is securing information during communication. Steganography can be defined as hiding information into another medium of information. Though cryptography has the same goal as steganography that is to keep the information secured, the main purpose of it makes a big different. With the purpose of opening an opportunity to create a communication channel that is covert and subliminal between two parties where by the existence of the particular message being sent and received are kept unknown or innocuous to any possible attacker.

Steganography can be divided into two that is technical steganography and linguistic steganography. There have been vigorous studies and researches in the technical aspect consisting images, audio and audio, but only few have manage to explore the potential use of linguistics, in other word is natural languages. Briefly, Linguistic steganography can be defined as the art of using written natural language to conceal secret messages. The whole idea is to hide the very existence of the real message. But it is not as much as studies are being done in technical steganography [4]. Those methods uses different kind of aspects and elements of the natural language like by replacing certain words with its synonyms. Spaces between words or white spaces are also being manipulated as a source of carrier to secret messages. The linguistic parts like the syntactic and semantic also have been studied and used as a medium to hide information.

DISCUSSION

Linguistic Stegabography Model

Linguistic Steganography is still in development and becoming more popular. This is because the usage of encryption standard it compatible with hiding in text method, in this case linguistic steganography. If the encrypted message is hidden in an image, not every pixel may be suitable for encoding ciphertext.

This is because changes to pixel might be visible in large fields of monochrome colour, or that lie on sharply defined boundaries [1]. There are several methods that have been developed to secure text-based document like marking a document using a codeword, manipulating semantic information to carry secret data and transforming ciphertext into an innocuous sentence by using a software system [6]. However, syntactic and semantic is constantly being used when dealing with natural language and has the potential to developed and used to hide information in documents [2].

Figure 1 shows the steganographic model for text that we assume. Text data that is used as a medium to embed secret data is called cover text. Cover text that contains secret data is called stego text. Secret data is embedded into cover text using a key. The sender sends the stego-text to the receiver through the Internet. The key the sender used is sent to the receiver through the private line. The receiver can extract embedded message from stego-text using the secret key [5].

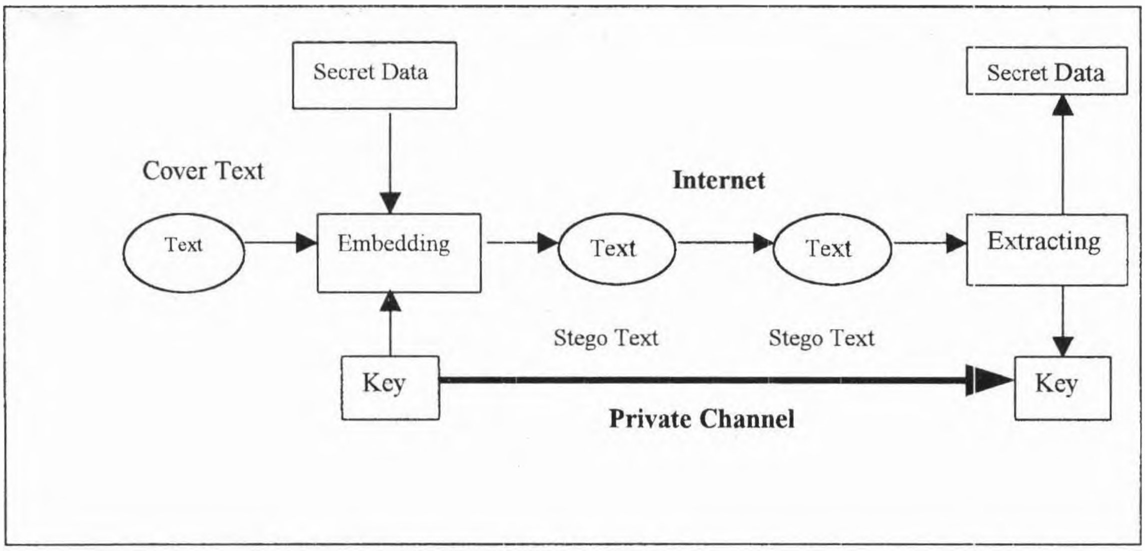

Figure 1. Model of a System for Linguistic Steganography [5]

Most of the researches done in linguistic steganography uses English as the main source of medium in linguistics. This paper will be looking beyond English and venture into other native languages. In this case, we have restricted our studies to three main languages uses in the Asia region, which are Chinese, Japanese, and Malay. The main focus of this paper is to discuss the potential of Malay language sentence structure and its morphological characteristic.

Language Comparison Study

Syntactic and semantic of a sentence is essential to proof the integrity of the sentence itself, especially a translated system. A comparison of main sentence and its direct words translation will allow us to analyse the syntactic and then see whether it is semantically correct. Figure 2 shows a direct word-to-word translation from Malay to English. If we check syntactically, there are minor errors in translating ***"telah mendapat"*** because of it can be represent by the word ***"got"***. On general basis, semantically the meaning of the direct word translated sentence is still understandable.

I ***got*** ***permission*** ***to*** ***take*** ***a*** ***vacation***

Saya telah <u>mendapat</u> keizinan untuk <u>mengambil</u> cuti rehat

Figure 2. Comparison Word-to-Word Translation of English Sentence to Malay

In Figure 3, only minor syntactic errors occur when translating Chinese sentence into English and Malay sentence. This is because the Chinese sentence pattern is similar to English (Subject + Verb + Object). The main idea of the sentence is also understandable even though there are several syntactic errors. Figure 4 is the comparison of a direct word translation from Japanese to English and Malay. Here we can see the main flaw in the translation. This occurs because the Japanese sentence pattern (Subject + Object + Verb) is different from English, Chinese and Malay. The syntactic error causes by disambiguation of the verb "*toru*", and "*tori*" which has the same meaning as "ambil" (take in English).

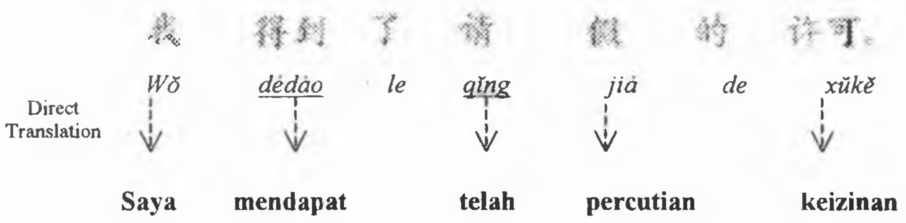

Figure 3: Comparison of Word-to-Word Translation of a Chinese Sentence to Malay

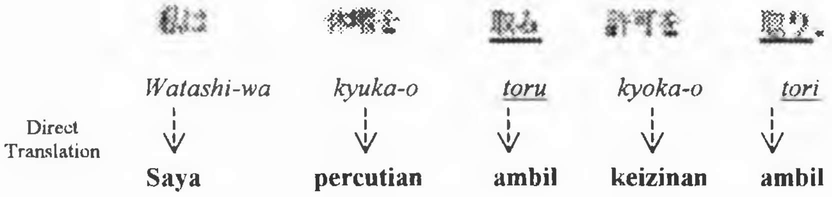

Figure 4: Comparison of Word-to-Word Translation of a Japanese to Malay

From the translation, we can see that in terms of syntactic of a sentence, Malay have much more in common with English compared to Japanese. This shows that Japanese syntactic is not significant and alteration is needed for a Japanese sentence to be translated in English. Differ from Malay; Japanese and Chinese has its own unique scripting, which have much more in common with English. When translating it into machine language or programming language Japanese, and Chinese and will have to define first their writing scripts according to the right orthography or spelling. Unlike English, Japanese language does not separate words with white spaces. The comparison may look very simple and straight forward, but the important aspect lies in the outcome of the word translation. From the above comparison, we can see that the English-Malay translation is still understandable compared to Chinese-Malay and Japanese-Malay translation.

Advantages of Malay Language

Every language has the capability to conceal information in it own unique structure. However, in such cases Malay language would be the preferred among the others though it had been claims to several translation issues. Positively, this problem occurred in machine translation system shows some possible contribution to develop linguistic steganography using Malay language. Having the same letters in the English alphabet is the main criteria that makes Malay a suitable candidate.

Moreover, Malay shares the same sentence pattern "*Subject + Verb + Objective*" (SVO) with English and morphologically makes it easy to develop a translation system. Therefore Malay language is replaceable in any kind of technology, mechanism, method or approach that have been developed fully in English. During the construction of the Malay index words, there are many spelling variations. Rather than create multiple entries with all information identical except the index, we allowed a single record to have multiple index forms, with the preferred form. Here it shows that there are synonyms that be generated and the use of it is essential in the approach used in NICETEXT [3].

Another related issue is that some words found in on-line resources are not found in the various dictionaries used as reference. There can be two reasons for this: the first is the word does not exist in the standard Malay vocabulary but is somehow *borrowed* from another language and the second is the word does exist but is rather recent and has not been incorporated into the printed dictionaries. The former is common, where many words found during checking are of Indonesian that have not been accepted as standard Malay but are used by Malay language speakers in Malaysia. The latter is more problematic as there is no way of knowing if the words have been standardized with out looking for them in a large up-to-date Malay corpus. In this issue, it shows that Malay language is evolving rapidly by the use of *borrowed* vocabulary from other languages. Therefore when generating a sentence, the electronic vocabulary will be massive.

Another issue that arose while compiling, checking and editing the lexicon was how many derivations need to be listed. It is inefficient to list semantically transparent regular derivational forms, although they may be needed in a bilingual dictionary if their translations are irregular. Malay is an agglutinative language, with many derivations arising through affixation. Affixation is highly productive, for example, the verb ***guna*** **(use)** in its root form generates many derivations via affixation. To add to the complexity, the affixation process in Malay allows a maximum of three layers for any root word, for example ***berkeseorangan*** **(to suffer from loneliness)** has undergone three layers of affixation.

Table 1: Affixation Process In Malay

LAYER	**AFFIXATION**	**WORD**	**ENGLISH TRANSLATION**
None	None	Orang (root)	**Person**
First	***Se-*** *(prefix)*	***S****eorang*	**Alone**
Second	***Ke-an*** *(circumfixation)*	***Ke****seorang****an***	**Loneliness**
Third	***Ber-*** *(prefix)*	***Ber****keseorangan*	**To suffer from loneliness**

Integrating two elements, steganography and natural language (Malay) is a task that is indeed a potential merge. After reviewing several foreign languages and its characteristic that can be used in Linguistic steganography, it shows that Malay has the utmost and better chance to be implemented. But it is not as simple as replacing it with English There are still few tasks and challenge that needed to be solved. With the advantage of having the same alphabet to English, research on the average use of alphabet in Malay could be manipulated in steganography. Malay also has several disadvantages of its own like solving lots of word generated by affixation and this problem occurred in computational translation system. But we can turn the problems around and use it as an advantage in using Malay in steganography. Based on this we have come out with the idea that the affixation could be used as the carrier for the semantic information of the embedded information [5].

Potential Steganography In ICT

Information and Communication Technology (ICT) has been growing rapidly with a lot of conventional applications under going total transition. Hence, it will surely be threaten by several security issues. Therefore, steganography could be applied in securing several critical aspects that will surely benefit each and every party that have direct and indirect involvement.

Covert Communications: These are mainly applications of steganography. In many situations, such as intelligence applications, people would like to send messages to each other without being detected. In such cases, the adversary is usually the enemy. It is often that the enemy does not detect the presence of a secret message transmitted.

Authentication: Sometimes it is necessary to verify the authenticity of input data, to determine whether the data at hand are original, fake, or some altered version of the original. In medical applications, for instance, it is of great significance to have the original data for diagnosis or treatment purposes. In that case, the data to be used may have been subject to some unintentional changes (e.g., compression, etc.) and should not be used under such conditions. In this scenario, it is also extremely important to verify the authenticity of the input data (often without knowledge of the original data). For authentication purposes, fragile watermarks seem to be a good solution; a properly designed fragile watermarking algorithm should be able to detect any change in the original data.

Identification and Proof: Such applications are usually targeted by robust watermarking algorithms and intended for commercial purposes. In such a situation, a company that produces and sells digital audio clips (e.g., Sony) or a movie company that sells its products over the Internet is concerned with copyright issues. In particular, it is very profitable for hackers to crack these products and sell them at a cheaper price. In such situations, original producers would like to have legally valid proof that they are the real owners. Robust signature casting is a possible solution in such cases. A practical precaution to overcome this problem could be the following: if a mutual agreement is reached between the producers of such digital goods and the producers of digital media players, then the unauthorized users can possibly be discouraged from unauthorized copying via robust watermarking.

Customer Tracing: Such applications are mainly intended for the fingerprinting problem. In such a situation, for example, a movie company inserts user IDs in each product before selling it. Whenever an unauthorized user is caught playing the movie or selling it, that user and his accomplices (the parties that were involved in producing that unauthorized copy) would be identified.

Data Embedding in Communications: In some communications applications, it is desirable to embed information into the host data before transmission. These are non-adversarial applications. For instance, within an in-band captioning scenario, it is desired to send the captions of a digital video via embedding the captions in the video signal, yielding little or no distortion visually, thereby reducing the required bandwidth for data transmission.

CONCLUSION

There are lots of natural languages being use all around the world. The growth is not just for the purpose of verbal communication but also as a tool of developing technologies. In this case, using it as a medium to carry secret messages in a document written in natural language is so that any possible attacker would have not expected it have embedded message and find it as innocuous. Japanese and Chinese have a massive amount of users in the Asian region and have the potential to be used as a medium in communicating and transferring information. But with the growing of usage in Malay and lots of researches have been done by foreign researchers, this definitely shows its capability to be used widely in many information technology areas, specifically linguistic steganography. With this study we hope to attain several overview on the linguistic aspect of languages so that further development in linguistic steganography can be made.

REFERENCES

1. Anderson, R. J. and Petitcolas, F. A. P. 1998. On The Limits of Steganography. *IEEE J. of Selected Areas in Comm.* 16(4): 474-481.

2. Bender, W., Gruhl, D., Lu, A., and Morimoto, N. 1996, Techniques for data hiding, *IBM Systems Journal*, 35, nos 3&4

3. Chapman, M., and Davida,G. 1997. Hiding the Hidden: A Software System for Concealing Ciphertext as Innocuous Text, ICICS'97. 335-343

4. Chapman, M., Davida, G. and Rennhard, M. 2001. A Practical and Effective Approach to Large-scale Automated Linguistic Steganography, ISC2001, LNCS2200, Springer-Verlag, Berlin Heidelberg. 156-165

5. Niimi, M., Minewaki, S., Noda, H., and Kawaguchi, E. 2003. Linguistic Steganography Using SD-Form Semantics Model, *IPSJ Journal*, 4(8): 1894-1903

6. Zolkafli, A, and Din, R. 2003. Linguistic Steganography: Securing Text based Document and Security Challenges. *National ICT Seminar 2003*. 122-131.