Research on Coverless Text Steganography Based on Single Bit Rules

Ning Wu1, Poli Shang2, Jin Fan3, Zhongliang Yang4, Weibo Ma1, Zhenru Liu1

1School of Electrical Engineering, Lanzhou Institute of Technology, Lanzhou 730050, China
2School of Electronic and Electrical Engineering, Lanzhou Petrochemical Polytechnic, Lanzhou 730060, China
3School of Information Science and Engineering, Lanzhou University, Lanzhou 730000, China
4Department of Electronic Engineering, Tsinghua University, Beijing 100084, China
wun0016@163.com

Abstract. Text steganography has become a hotspot in the field of information security in recent years. As an important steganographic carrier, text does not have the advantage that other multimedia could easily use more redundancy to hide information. Traditional methods were realized by changing the appearance attributes of text or by equivalent replacement based on grammar and semantics, but there were some shortcomings more or less. Based on Markov chain model, this paper proposed a coverless text steganography method. In order to enhance the effect of steganography, the method focused on the most important concepts in Markov model, the transition probability, and guided the generation of steganographic text based on it. The relevant experiments were designed to evaluate the model. The experimental results showed that the model had higher concealment and hiding capacity.

1. Introduction
Information hiding refers to the technical means of hiding meaningful information into non-secret carriers and transmitting it through open channels. As an important media, text is widely used. But not like other medias, text is less redundant space, so it is difficult to hide information. Accordingly, the research results of text-based information hiding technology are much less than others media's and its research significance is more prominent[1].

In recent years, the text information hiding method based on coverless text steganography has gradually attracted the attention of researchers. It could hide secret information by generating new text which accords with the statistical characteristics of natural language. It could better avoid steganalysis based on statistical features[2]. This paper proposed a coverless text steganography method based on single bit rules based on Markov chain model. In the method, the secret information was passed through the model, and then the corresponding steganographic texts were generated. In the process of text generation, the importance of transition probability was emphasized, and the transition probability was utilized to the greatest extent, aimed at generating steganographic text which was closer to the training text.
2. Related works

2.1. Markov chain
Markov chain, named Andre Markov, is a discrete stochastic process. It describes a sequence of states that, given current knowledge or information, the previous historical state of the process is invalid for predicting the future state. At each step of Markov chain, according to the probability distribution, the system can change from one state to another state or keep the current state.

2.2. Steganography based the model
The information hiding method based on coverless does not need true carriers, but generates encrypted text according to the statistical characteristics of natural language. Therefore, encrypted text can hardly be found by steganalysis based on statistical features of natural language, nor can it be found by comparing the differences between carrier text and encrypted text, which is securer[3].

[4] described a state transition diagram and the process of hiding and extracting information. Given a start state, the subsequent state transition process completed information hiding and extraction by matching with binary coding, as shown in Fig.1.

![State transition diagram and steganography](image)

Fig.1 State transition diagram and steganography

In the figure above, assuming that the information embedded in each state transition is 2 bits, when the hidden information needs to be embedded is 01 11, the encrypted text is S_0 S_2 S_7 according to the state transition diagram. The information extraction process is an inverse process of embedding process.

An important concept in state transition diagram is transition probability, which represents the probability of transition to the next specific state. In [5-6], some methods assumed that all transition probabilities from a given state to any other state were equal in order to simplify the generation process of natural text and related analysis. Other methods emphasized the importance of transition probability, but through analysis, the effect of text generation had not been greatly improved, and there was also the possibility of attack through statistical detection. In order to actively deal with steganalysis based on statistical features of natural language and to generate high quality natural sentences to the greatest extent, the steganography method was designed on the basis of Markov chain model, focusing on transition probability.

3. Coverless text steganography based on single bit rules

3.1. Establishment of state transition diagram
According to the selected sample library, all sentences with the same word as the beginning of the sentence are selected to prepare for the formation of the state transition diagram. Suppose that in a sample library, according to the selected statements, the established Markov chain state transition diagram is shown in Fig.2. In the diagram, the transition probability indicates the connection degree between the texts in the sample library.
3.2. Adjustment of state transition diagram, coding matching and steganography

For large sample texts, the state transition diagram formed has many branches, complex structure and high complexity of information matching and steganography, so the quality of texts generated will be affected. Therefore, at most two branches with the highest probability are reserved for each subsequent branch of the node in the state transition diagram, as shown in Fig.3.

In order to restore the steganography text to the correct original information, it is necessary to give a definite binary meaning to each branch of state transition based on the state transition diagram. Since there are at most two branches per node in the state transition diagram, the matching relationship of branches can be defined by only one binary number. For the state transition diagram shown in Fig.3, the matching results are shown in Fig.4.
When the secret information is input, the steganography texts can be generated according to the state transition diagram. Steganography algorithm can be expressed as follows:

1. Select the sample library and select all sentences with the same word as the beginning of the sentence.
2. According to the results of the first step, a Markov state transition diagram with the same word as the beginning of the sentence is established.
3. Adjust the state transition diagram to preserve at most two branches with the highest state transition probability for each node.
4. Binary code matches with the adjusted state transition diagram.
5. Text generation and steganography are accomplished according to the state transition diagram and input sequence.

3.3. Original information extraction

In order to restore original information correctly, the same state transition diagram should be used. Before the communication between the two sides, a unified sample library has been selected. A unified state transition diagram has been established, and it has been adjusted and matched with binary coding under the same rules. Therefore, when the receiver obtains the steganographic texts, the correct original information is extracted by comparing with the state transition diagram. The extraction algorithm of steganographic information can be expressed as follows:

1. Generate the state transition diagram according to the pre-agreed sample library.
2. Adjust the state transition diagram under the same rules.
3. Matching the binary information of the state transition diagram under the same rules.
4. Receive steganographic text from the sender.
5. According to the state transition diagram matching binary information, the steganographic text is compared with it, and the original information is obtained.

4. Experiments and results

In this section, the performance of the algorithm was tested through experiments. First, experiments datasets should be collected. Here, a collection of reviews about movies[7] and a collection about news[8] were chosen. The characteristics of the two datasets are as follows:

| Table 1 Test datasets |
|-----------------------|
| Dataset | Film review | News |
| Average Length | 19.94 | 22.24 |
| Sentence Number | 1,283,813 | 1,962,040 |
| Words Number | 25,601,794 | 43,626,829 |

In order to measure the performance of the algorithm, it would be tested from two different perspectives. First, as coverless text steganography algorithm, the primary purpose was to change the input information into new different text. Therefore, from the perspective of natural language processing and generation, the relevant indicators in this field could be used to measure the effect of steganography. Here, perplexity was chosen as an indicator. In natural language processing, the degree of perplexity could be used to measure the quality of language generation model. The smaller the perplexity was, the closer the statistical distribution of the generated text and the training text was. The perplexity is expressed by the following formula:

$$\text{Perplexity} = 2^{-\frac{\sum_{i=1}^{m} \log p(s_i)}}$$

where \(s_i = \{w_1, w_2, w_3, \ldots, w_m\}\) is the generated sentence, \(p(s_i)\) indicates the probability distribution over words in sentence \(s_i\), and the probability is calculated from the language model of the training texts. \(m\) is the total number of generated sentences. The Table 2 shows that our model's perplexity is smaller compared with the other models. This showed that the steganographic effect of our model was better.
Table 2: The perplexity of the algorithms

| Dataset | Baseline [7] | Baseline [8] | Ours   |
|---------|--------------|--------------|--------|
| IMDB    | 418.70±105.32| 161.92±143.31| 15.38±6.77 |
| News    | 470.54±122.73| 175.42±126.28| 17.05±15.21 |

Secondly, as a steganography algorithm, under the same conditions, more information, that was, the number of bits, was expected to be embedded into the text. Embedding Rate could reflect the performance of the algorithm. Its mathematical expression is as follows:

\[
ER = \frac{1}{N} \sum_{i=1}^{N} \frac{(L_i - 1)}{B(s_i)}
\]

where \(N\) is the number of generated sentences and \(L_i\) is the length of the \(i\)-th sentence. \(B(s_i)\) indicates the number of bits occupied by the \(i\)-th sentence in the computer. Two text steganography methods were compared with the method proposed. The test results are shown in Table 3.

Table 3: Embedding Rate comparison

| Methods                                 | Embedding Rate(%) |
|-----------------------------------------|-------------------|
| Method proposed in [9]                  | 0.33              |
| Method proposed in [10]                 | 1.0               |
| Ours                                    | 2.73              |

From the above table, the algorithm proposed in the paper could embed more information, that was, the steganography capacity of the algorithm was better.

5. Conclusion

This paper realized the steganography to protect information based on Markov model which was often used in natural language processing. For various reasons, previous related algorithms had not fully utilized or neglected the important concepts, transition probability, in Markov chains. In this paper, according to the model structure obtained from the samples, for each node, at most two branches with the highest transition probability were reserved. On this basis, information embedding and steganography were carried out in order to achieve better text steganography effect as far as possible. Through experiments, the performance of the method was proved from different aspects, and the effectiveness of the algorithm was verified.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (2018YFB1003205) and Gansu Natural Science Foundation (1506RJZA057).

References

[1] Luo Y, Huang Y. Text steganography with high embedding rate: Using recurrent neural networks to generate chinese classic poetry[C]//Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. ACM, 2017: 99-104.
[2] Sun, Huiyu, Ralph Grishman, and Yingchao Wang. Active learning based named entity recognition and its application in natural language coverless information hiding[J]. Journal of Internet Technology, 2017, 18(2): 443-451.
[3] Luo Y, Huang Y, Li F, et al. Text Steganography Based on Ci-poetry Generation Using Markov Chain Model[J]. KSII Transactions on Internet and Information Systems (TIIS), 2016, 10(9): 4568-4584.
[4] S. F. Wu, Researches on Information Hiding Technology[D]. Master Thesis. Hefei: China Science and Technology University, 2003 (in Chinese).
[5] Dai W, Yu Y, Dai Y, et al. Text Steganography System Using Markov Chain Source Model and DES Algorithm[J]. JSW, 2010, 5(7): 785-792.
[6] Moraldo H H. An Approach for text steganography based on Markov Chains[J]. arXiv preprint arXiv:1409.0915, 2014.
[7] Maas A L, Daly R E, Pham P T, et al. Learning word vectors for sentiment
analysis[C]/Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1. Association for Computational Linguistics, 2011: 142-150.

[8] Andrew Thomson. News dataset. DOI= https://www.kaggle.com/snapcrack/all-the-news/data

[9] Stutsman R, Grothoff C, Atallah M, et al. Lost in just the translation[C]/Proceedings of the 2006 ACM symposium on Applied computing. ACM, 2006: 338-345.

[10] Chen X, Sun H, Tobe Y, et al. Coverless information hiding method based on the chinese mathematical expression[C]/International Conference on Cloud Computing and Security. Springer, Cham, 2015: 133-143.