Coverless Text Steganography Based on Half Frequency Crossover Rule

Ning Wu
College of Software Engineering
Lanzhou Institute of Technology
Lanzhou, China
wun0016@163.com

Weibo Ma
School of Electrical Engineering
Lanzhou Institute of Technology
Lanzhou, China

Zhenru Liu
School of Electrical Engineering
Lanzhou Institute of Technology
Lanzhou, China

Poli Shang
School of Electronic and Electrical Engineering
Lanzhou Petrochemical Polytechnic
Lanzhou, China

Zhongliang Yang
Department of Electronic Engineering
Tsinghua University
Beijing, China

Jin Fan
School of Information Science and Engineering
Lanzhou University
Lanzhou, China

***Abstract*—In the development process of the information hiding technology, text steganography is a challenging task for researchers because of the small redundancy of information and the difficulty in hiding information. In this paper, a coverless text steganography method is proposed based on the Markov model and the half frequency crossover rule, which accords with the statistical characteristics of natural language. In order to prove the effectiveness of the algorithm, we select two large scale datasets for modeling and testing. The experimental results show that, compared with similar methods, the proposed model shows better concealment and hidden capacity in steganography.**

Keywords-coverless; text steganography; Markov chain; half frequency crossover

I. Introduction

The method that the sender hides the secret information to be transmitted in the host signal for transmission is called information hiding. As early as the Renaissance of Literature and Art, some people protected secret information by scattering them through the whole letter according to certain rules, which was also one of the ways of early information hiding. With the increase of Internet applications, data secure communication, identity authentication, intellectual property protection, and content recovery have become the focus of social attention [1]. The purpose of information hiding technology research is how to better protect important information. Text information is difficult to hide because of the lack of redundant space, so the research significance is more prominent. Based on the Markov model and half frequency crossover rule, this paper proposes a coverless text steganography method. The method tries to better preserve the natural language features of the training text and complete the generation of hidden text on the basis of more flexible linguistic relations [2].

II. Related Work

A. Coverless Text Steganography

Coverless text steganography is a hot research topic in the field of information hiding in recent years. Its advantages mainly lie in that the text generated by this method is more in line with the statistical characteristics of natural language. Therefore, it has better resistance to malicious attacks and stronger concealment.

Compared with the traditional information hiding technology, it only needs the communication parties to establish certain rules in advance and generates encrypted text directly according to the language model established from the training text, without carriers. Coverless text steganography has three characteristics: no embedding, no modification, and anti-steganalysis.

B. Markov Chain

Markov chain is a time-discrete, state-discrete and no-aftereffect stochastic process. That is to say, given the current knowledge or information, the current historical state of the process is ineffective in predicting the future state of the current. In the Markov chain, the system can change from one state to another or keep the current state. The transition probability determines this possibility.

The authors would like to thank the Natural Science Foundation Program of Gansu (1506RJZA057) and 2019 Longyuan Youth Innovative and Entrepreneurial Talents (Team) Project (Gan Zu Tong Zi [2019] No.39) (No.23) for supporting this research.

978-1-7281-4689-8/19/$31.00 ©2019 IEEE
DOI 10.1109/ICMCCE48743.2019.00168

III. Coverless Text Steganography Based on Half Frequency Crossover Rules

A. Sentence Generation Based on Markov State Transition Diagram

In the field of natural language processing, the statistical language model is widely used to model a text, and the Markov model could be used as a better approximation of the statistical language model. Through this model, it is easy to generate text similar to natural language. The state transition diagram is the core of the model. Given a start state, the subsequent state transitions are given binary information respectively. Information hiding and extraction are based on this binary information [3], as shown in Fig.1.

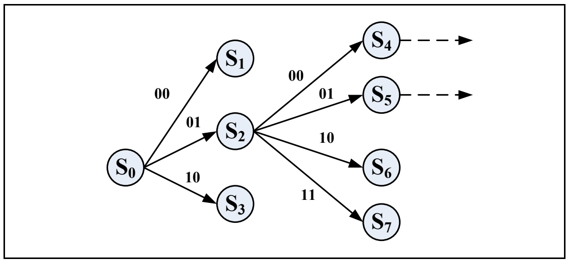

Figure 1. Information hiding based Markov chain

The start state is given as S_0. Supposing that the information embedded in each state transition is 2 bit. When the hidden information is 01 11, the text $S_0S_2S_7$ can be obtained according to the state transition diagram.

B. Half Frequency Crossover

First, the frequency of the initial text in the dataset is counted. This paper takes 3-gram and 4-gram half-frequency crossover as an example to model and test. For the whole dataset, a sequence of three words is selected, and the third word in each phrase is used as the candidate word of the first two words. The state transition diagram of the text model is established and every node is sorted from high to low according to its transition probability by using the method of [4]. Then a sequence including four words is selected and the last word as a candidate for the first three words is used. Then the state transition diagram under this rule is established in the same way.

After the secret information flow is ready, the encrypted text can be generated using the two language models that have been established. First, the initial words are randomly selected. Then the state transition diagram based on the 3-gram model is used to code. If the number of bits embedded in each candidate word is n, the first 2^n words with the highest transition probability in the candidate pool are selected by the method of [5]. These words match the n-bit binary encoding information corresponding to the $0 \sim 2^n$-1 decimal number, respectively. n bits are read in and steganographic text is generated. When the number of words in a sentence is greater than 4, the next candidate words need to be generated by taking into account the 3-gram and 4-gram language models.

The specific method is to reduce the frequency of all candidate words in the 3-gram language model by half. If a candidate word in the last two words candidate pool exists in the last three words candidate pool, then adds its frequency in the last two words candidate pool to 1/2 of the frequency in the last three words candidate pool. Then, in the last two candidate pools, all candidate words are sorted from large to small according to the recalculated transition probability and coded according to the same coding rules. Then, the n bits to be embedded are read backward to generate matching coded candidate words. If there is no candidate word with the same encoding as bit streams to be embedded in the candidate pool, the sentence will be generated and the initial word will be selected randomly to repeat the embedding process until all the information is embedded. Fig.2 further illustrates the process of reordering candidates in the method.

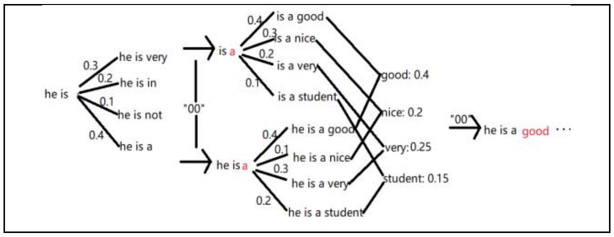

Figure 2. Half frequency crossover

Once the sentence length satisfies the state transition of 4-gram, the probability of the current candidate words needs to be recalculated and sorted. The reordered lexicons in the figure are "good", "very", "nice", "student" and their codes will be "00", "01", "10", "11". Assuming that the hidden bitstream is "0000...", the sentence is generated as "he is good...".

The algorithm of generating steganography text is as follows.

Algorithm 1. Text Steganographic

```
Input:
Training dataset D
Secret bit stream B= {$b_1$, $b_2$, ..., $b_m$}
Key list K = {$key_1$, $key_2$, ..., $key_n$}
Output:
    The steganography text S= {$s_1$, $s_2$, ..., $s_p$}

Establish a state transition diagram and adjust it;
while not the end of B do
    Randomly select the initial phrases of the generated sentence;
    Match the binary code according to the candidate word size of
  the last words based on preceding gram language model;
    Read n bitstream;
    Generate the matching words in the candidate words;
    if the current words of the candidate are also in the latter gram
  language model's candidate words pool
    Half frequency crossover adjustment;
    Reorder candidate word frequency after recalculation;
return steganography text S
```

C. Information Extraction

In order to extract the original secret information accurately, the receiver needs to select the same corpus and build the same state transition diagram as the sender. Then the state transition diagram is adjusted with the same rules and the final model is formed. Both sides also agreed to match the binary code with the model in a consistent manner. Only in this way can the receiver receive the steganographic text and extract the original information correctly by the process opposite to information hiding. The information extraction algorithm is as follows.

Algorithm 2 Information Extraction

```
Input:
The steganography text S={$s_1$,$s_2$, ..., $s_p$}
Output:
Secret bit stream B = {$b_1$, $b_2$, ..., $b_m$}

Establish the same state transition diagram as sender and adjust it with the same rules;
while not the end of the text do
    if not the end of the current sentence
        Read the initial phrases of the sentence and find the corresponding state transition diagram;
    else if the current words of candidate word k based on preceding gram language model also exist in the latter gram language model's candidate pool
        Half frequency crossover adjustment;
        Extract the binary code of the last word;
    else
        Read the first two words of the next sentence and find the corresponding state transition diagram;
return secret bit stream B
```

IV. Experiments and Analysis

In natural language generation technology, the purpose of any language model is to generate text that can meet the characteristics of natural language. This is the significance of the research and implementation of text information hiding technology. Based on the language model proposed in this paper, we choose two datasets to train and test the model, which are IMDB [6] and news [7]. These datasets contain a large number of texts with different language characteristics, such as professional vocabulary, daily vocabulary, popular vocabulary, etc. Experiments based on these datasets could more objectively verify the effectiveness of the proposed method. The characteristics of the two datasets are as follows.

TABLE I. Test Datasets

Dataset	IMDB	News
Average Length	19.94	22.24
Sentence Number	1,283,813	1,962,040
Words Number	25,601,794	43,626,829

In information hiding technology, the quality of the generated text is described by concealment. The closer the features of generated text and training text are, the better the anti-detection performance of generated text in the process of public channel propagation and the higher the concealment. Indicators of natural language processing technology can be used to measure the concealment of language models. Perplexity is such an indicator. There is a negative correlation between the degree of perplexity and the quality of the generated sentences. Its expression is as follows.

$$Perplexity = 2^{-\frac{1}{n}\sum_{i=1}^{n} logp(s_i)} \tag{1}$$

where $s_i=\{w_1, w_2, w_3,...,w_n\}$ is the generated sentence, n is the total number of generated sentences. $p(s_i)$ is the probability of a sentence. The calculation of $p(s_i)$ depends on the n-gram model, that is, formula (2).

$$p(w_1,w_2,\ldots,w_\mathrm{n}) \\ =p(w_1)p(w_2|w_1)p(w_3|w_1w_2)\ldots p(w_\mathrm{n}|w_1,\ldots,w_{n\text{-}1}) \tag{2}$$

The TABLEII shows that our model's perplexity is smaller compared with the other similar models. It means that the steganographic effect of our model is better.

TABLE II. The Perplexity of The Algorithms

Dataset	Baseline [8]	Baseline [9]	Ours
IMDB	418.70±105.32	161.92±143.31	15.97±7.57
News	470.54±122.73	175.42±126.28	17.41±8.91

Hidden capacity is usually described by the embedding rate. In the field of information hiding, for most algorithms, it is difficult to ensure that the concealment is improved while maintaining a high embedding rate. In other words, between concealment and embedding capacity, only one of the indicators can be satisfactorily met. The embedding rate is calculated by the formula (3).

$$ER = \frac{1}{N}\sum_{i=1}^{\mathrm{N}}\frac{(L_i\text{-}1)}{B(s_i)} \tag{3}$$

where N is the number of generated sentences and L_i is the length of *i-th* sentence. $B(s_i)$ indicates the number of bits occupied by the *i-th* sentence on the computer. The test results are shown in TABLE III.

TABLE III. Embedding Rate Comparison

Methods	Embedding Rate (%)
Method proposed in [10]	0.33
Method proposed in [11]	1.0
Ours	2.78

As can be seen from TABLE III, the proposed algorithm achieves more information embedding. Therefore, the hidden capacity of the algorithm is better.

V. Conclusion

In this paper, we try to combine different language rules with training texts in the half frequency crossover way based on the Markov model. It is hoped that the linguistic features embodied in different rules can be better displayed. Through the modeling and testing of two datasets with different features, it is proved that the method not only achieves better hiding but also meets the requirement of larger capacity, which reduces the possibility of detection when the generated text propagates in public channels.

References

[1] Luo, Yubo, and Y. Huang. "Text Steganography with High Embedding Rate: Using Recurrent Neural Networks to Generate Chinese Classic Poetry." *the 5th ACM Workshop*ACM, 2017.

[2] Luo, Yubo, et al. "Text Steganography Based on Ci-poetry Generation Using Markov Chain Model." *TIIS* 10.9 (2016): 4568-4584.

[3] S. F. Wu."Researches on Information Hiding Technology". Master Thesis. Hefei: China Science and Technology University, 2003(in Chinese).

[4] Wu, Ning, et al. "Research on Coverless Text Steganography Based on Single Bit Rules." *Journal of Physics: Conference Series*. Vol. 1237. No. 2. IOP Publishing, 2019.

[5] Wu, Ning, et al. "Coverless Text Steganography Based on Maximum

Variable Bit Embedding Rules." *Journal of Physics: Conference Series*. Vol. 1237. No. 2. IOP Publishing, 2019.

[6] Maas, Andrew L., et al. "Learning word vectors for sentiment analysis." *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. Association for Computational Linguistics, 2011.

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

[8] Dai, Weihui, et al. "Text Steganography System Using Markov Chain Source Model and DES Algorithm." JSW 5.7 (2010): 785-792.

[9] Moraldo, H. Hernan. "An Approach for text steganography based on Markov Chains." *arXiv preprint arXiv:1409.0915*(2014).

[10] Stutsman, Ryan, et al. "Lost in just the translation." *Proceedings of the 2006 ACM symposium on Applied computing*. ACM, 2006.

[11] Chen, Xianyi, et al. "Coverless information hiding method based on the Chinese mathematical expression." *International Conference on Cloud Computing and Security*. Springer, Cham, 2015.