A novel linguistic steganalysis method for hybrid steganographic texts

Yimin Xu¹, Liran Yang¹, Tengyun Zhao¹ and Ping Zhong²*

¹College of Information and Electrical Engineering, China Agricultural University, Beijing, 100083, China
²College of Science, China Agricultural University, Beijing, 100083, China

*Corresponding author’s e-mail: zping@cau.edu.cn

Abstract. Most of the existing linguistic steganalysis methods mainly focus on detecting steganographic texts which are generated by embedding secret information into a type of text medium using one steganographic algorithm. But in practical applications, a large number of the steganographic texts may be hybrid ones which are generated by embedding secret information into different types of text media using different steganographic algorithms. In this paper, inspired by transfer learning, a novel linguistic steganalysis method is proposed to detect hybrid steganographic texts. The proposed method first uses the pre-trained BERT language model to obtain initial context-dependent word representations. Then the extracted features are fed into attentional Long Short-Term Memory (LSTM) to get the final contextual representations of sentences. The experimental results show that the proposed method can better satisfy the practical application demands than the existing linguistic steganalysis methods.

1. Introduction

Steganography is the science of embedding secret information into common carriers such as text [1-2], image [3] and audio [4]. Due to the frequent use of text in daily life, text steganography has attracted extensive attention. The text steganography can be mainly divided into two categories: modification-based steganography [6-9] and generation-based steganography [10-12]. The former embeds secret information mainly by modifying cover texts and the latter usually automatically generates steganographic texts according to the secret information.

As a counter technique of text steganography, linguistic steganalysis aims to detect whether the texts contain secret information. Most of the traditional linguistic steganalysis methods are based on machine learning technology [13-15]. These methods mainly use manually extracted features, which makes the extracted features simple and not universal. Recently, some linguistic steganalysis methods based on deep learning technology have been proposed [16-20]. For example, Wen et al. [18] utilized convolutional neural network (CNN) to extract local correlation features of different granularities. Yang et al. [19] employed the Bidirectional Long Short-Term Memory (Bi-LSTM) to extract long-distance contextual dependency information. Niu et al. [20] combined the advantages of Bi-LSTM and CNN to extract both the global semantic features and local features.

Although the existing linguistic steganalysis methods based on deep learning technology have made great progress compared with the traditional ones, most of them mainly focus on detecting steganographic texts which are generated by embedding secret information into a type of text medium using one steganographic algorithm. While in practical applications, the bulk of the steganographic texts
may be hybrid steganographic texts, namely, these steganographic texts are generated by embedding secret information into different types of text media using different steganographic algorithms.

In the light of the above point, a novel linguistic steganalysis method is proposed in this paper. Firstly, motivated by transfer learning, BERT [21] is introduced to obtain initial context-dependent word representations. By using BERT, the proposed method can extract the correlations between words from different representation subspaces. Furthermore, Bi-LSTM and attention mechanism are adopted to fuse the contextual information and get the final representations of sentences. The experimental results demonstrate the advantages of the proposed model in meeting the application requirements.

The rest of the paper is organized as follows. Section 2 introduces the proposed method in detail. The next section presents the experimental settings and analyzes the experimental results. Finally, the conclusions are summarized in Section 4.

2. The proposed method

The proposed method can mainly split into three parts: (1) obtaining initial word representations; (2) fusing contextual information and getting the final representations of sentences. (3) forecasting whether the input sentences contain secret information. For clarity, the overall structure of the proposed method is shown in figure 1.

![Diagram of the proposed method](image)

For linguistic steganalysis, the main criteria for judging whether the texts contain secret information are the differences in semantic spatial distribution between cover and stego texts. Usually, the correlations between words in a sentence will be destroyed after embedding secret information, which will lead to the changes in semantic spatial distribution. The main components of BERT are transformer encoders [22] which mainly depend on multi-head self-attention. For each input sentence, multi-head self-attention can directly capture the correlations between words from different representation subspaces, which is very helpful to improve the performance of linguistic steganalysis.

In addition, the conditional probability in sentences also plays an important role in linguistic steganalysis. We assume a sentence $S$ of length $n$, that is, $S = \{x_1, x_2, \cdots, x_n\}$, where $x_i$ indicates the $i$-th word in $S$. The conditional probability of $S$ can be expressed as:

$$p(S) = p(x_1, x_2, \cdots, x_n) = p(x_1)p(x_2 | x_1) \cdots p(x_n | x_1, x_2, \cdots, x_{n-1})$$  (1)

Once a sentence is embedded with secret information, the probability distribution of the entire sentence will be inevitably affected. As a variant of RNN, LSTM can not only grab the changes of the semantic space caused by the destruction of conditional probability of texts, but also alleviate gradient vanish problem. After Bi-LSTM, attention mechanism is applied to automatically capture important clues to linguistic steganalysis.

Finally, the fully connected layer with dropout is used to prevent overfitting and soft-max function tries to judge whether the input sentences are embedded with secret information.
3. Experiments

3.1. Experimental datasets and settings

In the experiments, we construct multiple hybrid steganographic text datasets by using two steganographic algorithms: Tian-Fang [1] and T-Lex [2], which are very classic and widely used ones. Tian-Fang is a generation-based steganographic algorithm which can generate steganographic texts with different embedding rates. T-Lex is a modification-based steganographic algorithm which embeds secret information into texts based on synonym substitution. We use Tian-Fang to embed secret bitstream into three common types of texts media (IMDB [23], News [24] and Twitter [25]), respectively. Similarly, we use T-Lex to embed secret bitstream into these three types of text media, respectively. Then the steganographic texts generated by the two steganographic algorithms are used to construct the hybrid steganographic text datasets in different ways, which are introduced in the following corresponding subsections. Each dataset is divided into training set, validation set and test set according to the ratio of 72:8:20. Moreover, three representative linguistic steganalysis methods, including LS-CNN [18], TS-BIRNN [19] and R_BILSTM_C [20], are chosen as our comparison methods. Accuracy(Acc), precision(P) and recall(R) are used as the evaluation indicators.

For the pre-trained BERT model, we choose the “BERTBase-uncased” version which contains 12 Transformer encoder blocks. The hidden size is 768 and the number of self-attention heads is 12. Some hyper-parameters of the proposed method are set as follows: The hidden size in Bi-LSTM is 256, the nonlinear activation function is GELU and the optimizer is AdamW. The batch size is 16, the dropout rate is 0.5, and the initial learning rate is 5e-5.

3.2. Experimental results

(1) The first experiment is to verify the performance of the proposed method when text media are hybrid. Specifically, we mix the steganographic texts of various text media generated by Tian-Fang with bpw=1,2,3,4,5, respectively. And mix the steganographic texts of various text media generated by the T-Lex. Then we get six datasets, and each contains three types of text media and 6000 cover-stego text pairs. The experimental results are shown in Table 1, where the “TF” denotes Tian-Fang steganographic algorithm.

(2) The second experiment is to evaluate the performance of the proposed method when embedding rates are hybrid. Specifically, for each medium, the steganographic texts generated by Tian-Fang at various embedding rates (bpw=1,2,3,4,5) are mixed. Then we get three datasets, and each one contains 10000 cover-stego text pairs. The experimental results are shown in Table 2.

(3) The third experiment is to test the performance of the proposed method when steganographic algorithms are hybrid. Specifically, for each medium, we mix the steganographic texts generated by Tian-Fang at various embedding rates (bpw=1,2,3,4,5) and the steganographic texts generated by T-Lex. Then we get three datasets, and each one contains one type of text medium, two steganographic algorithms and total 20000 of cover-stego text pairs. The experimental results are shown in Table 3.

Table 1. The detection results when text media are hybrid.

| Methods   | LS-CNN [18] | TS-BIRNN [19] | R_BILSTM_C [20] | Proposed |
|-----------|-------------|---------------|-----------------|----------|
| Steganography | Acc | P | R | Acc | P | R | Acc | P | R | Acc | P | R |
| TF(bpw=1)  | 0.777 | 0.765 | **0.800** | 0.768 | 0.784 | 0.741 | 0.768 | 0.755 | 0.795 | **0.804** | 0.839 | 0.752 |
| TF(bpw=2)  | 0.835 | 0.825 | 0.852 | 0.833 | 0.822 | 0.849 | 0.843 | 0.842 | 0.846 | **0.885** | 0.904 | 0.862 |
| TF(bpw=3)  | 0.909 | 0.915 | 0.902 | 0.885 | 0.859 | **0.921** | 0.904 | 0.912 | 0.895 | **0.933** | 0.958 | 0.905 |
| TF(bpw=4)  | 0.954 | 0.955 | **0.953** | 0.937 | 0.935 | 0.940 | 0.945 | 0.941 | 0.948 | **0.963** | 0.974 | 0.952 |
| TF(bpw=5)  | 0.957 | 0.971 | 0.943 | 0.943 | 0.951 | 0.934 | 0.953 | 0.957 | 0.948 | **0.978** | 0.986 | **0.962** |
| T-Lex      | 0.839 | **0.919** | 0.743 | 0.827 | 0.894 | 0.741 | 0.830 | 0.904 | 0.738 | **0.858** | 0.902 | **0.802** |

Table 2. The detection results when embedding rates are hybrid.

| Methods | LS-CNN [18] | TS-BIRNN [19] | R_BILSTM_C [20] | Proposed |
|---------|-------------|---------------|-----------------|----------|
| Text media | Acc | P | R | Acc | P | R | Acc | P | R | Acc | P | R |
| News    | 0.966 | 0.949 | **0.985** | 0.960 | 0.940 | 0.983 | 0.961 | 0.952 | 0.970 | **0.971** | 0.970 | 0.973 |
3.3. Experimental analysis

Firstly, from the above experiments, it can be observed that the proposed method achieves state-of-the-art performance in various hybrid scenes, including hybrid text media, hybrid embedding rates and hybrid steganographic algorithms. This demonstrates that no matter what steganographic algorithm is used to embed secret information into what type of text media, the features extracted by the proposed method are sufficiently comprehensive to expose the distribution changes in the semantic space before and after steganography, so that the proposed method can accurately distinguish the stego texts from a large number of normal-looking texts.

Besides, it can be found from Table 2 that the detection performance of each method for long texts like News and IMDB is better than that for short texts like Twitter. This is because if the cover texts are short, the steganographic algorithm can generate stego texts with high naturalness, which makes the semantic spatial distribution between the cover and stego texts more similar. Therefore, the cover and stego texts are more difficult to be distinguished.

4. Conclusion

Motivated by transfer learning, a novel linguistic steganalysis method has been proposed in this paper for hybrid steganographic texts. Multiple hybrid steganographic text datasets are constructed to verify the performance of the proposed method. The experimental results show that the proposed method can achieve state-of-the-art performance, which proves that the proposed method can be commendably applied to practice. In future work, we will explore the performance of linguistic steganalysis method when the text media, embedding rates and steganographic algorithms are all mixed together.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 61872368.

References

[1] Weinstein, K. (1998) Lexical steganography through adaptive modulation of the word choice hash. http://web.mit.edu/keithw/tlex/.
[2] Fang, T., Jaggi, M., Argyraki, K. (2017) Generating steganographic text with lstms. In: ACL 2017-55th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop. Vancouver. pp. 100–106.
[3] Wang, Y., Zhang, W., Li, W., Yu, X., Yu, N. (2020) Non-additive cost functions for color image steganography based on inter-channel correlations and differences. IEEE Transactions on Information Forensics and Security., 15: 2081–2095.
[4] Yi, X., Yang, K., Zhao, X., Wang, Y., Yu, H. (2019) Ahcm: Adaptive huffman code mapping for audio steganography based on psychoacoustic model. IEEE Transactions on Information Forensics and Security., 14: 2217–2231.
[5] Mahato, S., Khan, D.A., Yadav, D.K. (2020) A modified approach to data hiding in Microsoft word documents by change-tracking technique. Journal of King Saud University - Computer and Information Sciences., 32: 216 – 224.
[6] Xiang, L., Wang, X., Yang, C., Liu, P. (2017) A novel linguistic steganography based on synonym run-length encoding. IEICE transactions on Information and Systems., 100: 313–322.
[7] Hu, H., Zuo, X., Zhang, W., Yu, N. (2017) Adaptive text steganography by exploring statistical and linguistical distortion. In: 2017 IEEE Second International Conference on Data Science in Cyberspace (DSC). Shenzhen. pp. 145–150.
[8] Li, M., Mu, K., Zhong, P., Wen, J., Xue, Y. (2019) Generating steganographic image description by dynamic synonym substitution. Signal Processing., 164: 193–201.
[9] Huo, L., Xiao, Y. (2016) Synonym substitution-based steganographic algorithm with vector distance of two-gram dependency collocations. In: 2016 2nd IEEE International Conference on Computer and Communications (ICCC). Chengdu. pp. 2776–2780.
[10] Luo, Y., Huang, Y. (2017) Text steganography with high embedding rate: using recurrent neural networks to generate chinese classic poetry. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. Philadelphia. pp. 99–104.
[11] Wen, J., Zhou, X., Li, M., Zhong, P., Xue, Y. (2019) A novel natural language steganographic framework based on image description neural network. Journal of Visual Communication and Image Representation., 61: 157–169.
[12] Yang, Z., Guo, X., Chen, Z., Huang, Y., Zhang, Y.J. (2019) Rnn-stega: Linguistic steganography based on recurrent neural networks. IEEE Transactions on Information Forensics and Security., 14: 1280–1295.
[13] Chen, Z., Huang, L., Miao, H., Yang, W., Meng, P. (2011) Steganalysis against substitution-based linguistic steganography based on context clusters. Computers & Electrical Engineering., 37: 1071–1081.
[14] Chen, Z., Huang, L., Yang, W. (2011) Detection of substitution-based linguistic steganography by relative frequency analysis. Digital investigation., 8: 68–77.
[15] Taskiran, C.M., Topkara, U., Topkara, M., Delp, E.J. (2006) Attacks on lexical natural language steganography systems. In: Proceedings of SPIE - The International Society for Optical Engineering. San Jose. pp. 607209 – 607209–9.
[16] Xu, Y. (2020) Unsupervised Deep Learning for Text Steganalysis. In: 2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI). Qingdao. pp. 112-115.
[17] Bao, Y., Yang, H., Yang, Z., Liu, S., Huang, Y. (2020) Text steganalysis with attentional lstm-cnn. In: 2020 5th International Conference on Computer and communication Systems(ICCCS). Shanghai. pp. 138 – 142.
[18] Wen, J., Zhou, X., Zhong, P., Xue, Y. (2019) Convolutional neural network based text steganalysis. IEEE Signal Processing Letters., 26: 460–464.
[19] Yang, Z., Wang, K., Li, J., Huang, Y., Zhang, Y.J. (2019) Ts-rnn: Text steganalysis based on recurrent neural networks. IEEE Signal Processing Letters., 26: 1743–1747.
[20] Niu, Y., Wen, J., Zhong, P., Xue, Y. (2019) A hybrid r-bilstm-c neural network based text steganalysis. IEEE Signal Processing Letters., 26: 1907–1911.
[21] Devlin, J., Chang, M.W., Lee, K., Toutanova, K. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[22] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.,Polosukhin, I. (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach. pp. 5998 – 6008.
[23] Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C. (2011) Learning word vectors for sentiment analysis. In: Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies. Portland. pp. 142–150.
[24] Thompson, A. (2017) Kaggle. https://www.kaggle.com/snapcrack/all-the-news/data.
[25] Go, A., Bhayani, R., Huang, L. (2009) Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford. 1(12).