Detection Techniques for Chinese Jargon: A Survey

Lu Zhang, *Yeonjoon Lee
Major in Bio Artificial Intelligence, Dept. of Applied Artificial Intelligence, Hanyang University,
*Dept. of Computer Science and Engineering, Hanyang university
kitsch@hanyang.ac.kr, yeonjoonlee@hanyang.ac.kr

중국어 Jargon의 탐지 기술 연구 동향 분석

쟝루, *이연준
한양대학교 인공지능융합학과 바이오인공지능융합전공, *한양대학교 컴퓨터공학과

Abstract

Contents on the Internet are invaluable for understanding cybercrime. On the internet, criminals substitute jargon for authentic terms to avoid censorship and discovery. Obfuscated language can avoid automatic and even manual detection because of its harmless appearance, but it must be comprehended by the intended audience. There has been a significant amount of research on English jargon, but little on Chinese jargon. We offer an overview of recent automatic Chinese jargon detection methods in this study.

I Introduction

In recent years, as the online market for promotion has grown rapidly, the Internet is flooded with illicit promotional content. Online platforms and censorship have their own rules and detection methods for avoiding such content. When miscreants want to post their illicit content on the internet, they often use jargon that has an innocent or vague look to bypass such detections and censorship, and deliver information to target consumers. For example, "coke" may not be a kind of soft drink, but stand for "cocaine", "飛行 (flying)" does not mean the sport but stands for "吸毒 (take methamphetamine)." The manual identification of jargon requires professional knowledge to do fact-checking on suspicious remarks, which takes much time and human resources, so it is necessary to develop techniques to handle such content automatically. In the past, technologies based on feature engineering and machine learning were widely used. Because jargon is constantly updated and iterated, it means that whether it is manual review or automatic review of traditional technologies, knowledge needs to be constantly updated. Recently, with the growth of Deep Learning (DL) and Natural language processing (NLP), DL and NLP models are widely used in such scenarios. Compared to conventional techniques, they no longer rely on feature engineering but prefer to build models for feature learning and extract high-level representations of data, thus achieving better classification performance.

To date, the research on English jargon detection is relatively comprehensive and achieves significant results, whereas Chinese jargon is much less investigated. English jargon detection techniques cannot be directly applied to Chinese because of the substantial differences between Chinese and English, so it is necessary to develop Chinese jargon detection techniques.

II Method

In this section, we briefly introduce some methods used for Chinese jargon detection.

II.1 Machine Learning-based Method

To the best of our knowledge, Zhao et al. [1] proposed the earliest Chinese underground jargon automatic detection with two typical language models: word embedding and Latent Dirichlet Allocation (LDA) [2]. Zhao et al. collect data from the chat histories of QQ (the most widely used instant messaging platform in China) groups via keyword searching and snowball collecting, clean the obtained data, segment Chinese words using the Chinese Lexical Analysis System, and utilize word embedding and LDA. For word embedding, Zhao et al. interpret the semantics of the jargon by examining the closest words in the word embedding vector space, calculate the distance between them with cosine similarity. Because terms with comparable meanings have similar probability weights within the same topic, Zhao et al. interpret the semantics of the jargon by the key topics to which it belongs. As the earliest research on Chinese jargon automatic detection, this paper introduced NLP methods to deal with such problems and inspired subsequent research.

Table 1: Comparison of methods mentioned in this paper

Scenario	Method	Data Source
Chinese underground market	Word-embedding&LDA [1]	QQ group
Black Hat SEO campaign	Similarity of characters' pronunciation and shape [3]	Search engine (seed illicit keywords + related search)
Chinese darknet trading	BERT-based unsupervised learning [7]	Chinese darknet trading website

II.2 Similarity-based Method

Yang et al. [3]'s research focuses on the jargon for Black Hat SEO campaign and proposed an industrial-strength, scalable monitoring/detection solution for promotional website defacements called DMOS. A part of this work is jargon normalization. This work divides jargon into two categories: homophonic jargon, which stands for jargon with similar pronunciation to the original words, and homomorphic jargon, which stands for jargons with similar shape to the original words. Yang et al. first collect seed jargons through manual crafting and keyword extraction from illicit websites and cybercrime marketplaces. By related search of search engines, related words are also added to the seed jargon list after checking manually. Then, they use a language model to filter out smooth sentences because sentences with jargon should not be smooth, and use a sliding window to traverse the rest of the sentences to select characters. For homophonic jargons, they compare the distance of Pinyin (pronunciation) between the characters of unsmooth words and seed jargons; if the distance is below the threshold, it means the characters are suspicious jargons. For homomorphic jargons, Yang et al. use the Four-Cournel System [4] to compare the shapes of the characters and seed jargons to identify suspicious jargons. Experiments demonstrate DMOS performs better than Tag-aware Hierarchical Attention Network [5] and BERT [6] on jargon detection.

II.3 DL-based Method

Liang et al. [7] proposed an unsupervised detection framework called DC-BERT (Dark Corpus Bidirectional Encoder Representations from Transformers) for detecting Chinese jargon in the darknet. It has a detection accuracy of 91.5%, which is the highest rate of Chinese jargon detection to date, and classifies the jargons into 10 categories. They first build a crawler to collect data from the six most popular underground Chinese underground e-commerce websites, then clean and preprocess the collected data. Unlike other work uses context to determine jargon only, this work determines jargons by using both context and the Chinese dictionary. A word is determined to be jargon if it is not found in the Chinese dictionary or is found there but has a distant relative with the supposed context.

III Conclusion

This paper briefly introduces recent research on Chinese jargon detection and summarises them in Table 1. Although there hasn't been much research on Chinese jargon recognition, it has made tremendous progress. Different technologies are also suitable for different scenarios, such as the underground black market, Black Hat SEO, etc. With the widespread use of large language models, the development of this aspect will be more rapid.

ACKNOWLEDGMENT

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00155885, Artificial Intelligence Convergence Innovation Human Resources Development (Hanyang University ERICA)) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-01343, Artificial Intelligence Convergence Research Center(Hanyang University ERICA)).

References

[1] Zhao, Kangzhi, et al. "Chinese underground market jargon analysis based on unsupervised learning." 2016 IEEE Conference on Intelligence and Security Informatics (ISI). IEEE, 2016.

[2] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3.Jan (2003): 993-1022.

[3] Yang, Ronghai, et al. "Scalable Detection of Promotional Website Defacements in Black Hat SEO Campaigns." 30th USENIX Security Symposium (USENIX Security 21). 2021.

[4] Four-corner system, 1995.

[5] Yang, Zichao, et al. "Hierarchical attention networks for document classification." Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016.

[6] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[7] Ke, Liang, Xinyu Chen, and Haizhou Wang. "An Unsupervised Detection Framework for Chinese Jargons in the Darknet." Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. 2022.