Chinese Language under Censorship

Shifei Chen
5LN710, Natural Languages Processing, Uppsala University

1 Introduction

In 2017, 64% of netizens lived in countries where their freedom of expression is supressed and for the third consecutive year, China was the world's worst abuser of internet freedom [1]. The censorship system in China behaves in several ways including limited access to websites and keywords, removal of posts on social network websites and blogs, cyber attacks on websites and createing contents for public opinions manipulation. The system works behind th curtain so its organization and mechanism are still not clear to the public. Censored contents usually contain information about pornography, violence and last but least, information against the communist regime. But this doesn't mean all negative information will be censored eventually, there are plenty of them that are considered to be "safe" being expressed freely on the internet. Also, some of the contents that are obfuscated intentially can escape. To understand what kind of expression are more likely to be supressed from a linguistic prospective and how to avoid it by encoding languages, I have conducted this survey base on the First Workshop on Natural Language Processing for Internet Freedom (Brew et al., 2018).

2 Characteristics of Sensitive Words Blacklists

Censorship in China isn't implemented by the governments or its subordinaries. Service providers are required to comply with content regulations by themselves, which means tech companies and individual developers are liable for the content generated by their users. A simple solution is filtering unfavored information by using a balcklist, which isn't a novelty to block out contents and it is being widely carried out by tech companies and governments around the world, such as Google [2] and IWF [3].

Reverse engineering revealed some blacklists from popular Chinese mobile phone applicaitons (Knockel et al., 2011) (Knockel et al.,) and games (Knockel et al., 2017). These studies suggest that, "rather than Chinese censorship being top-down and monolithic, it is a decentralized system where developers are liable for deciding what to censor themselves" (Knockel et al., 2018) since the blacklists found have very little overlap between each other.

If it can be inferred that these big companies have compiled their own blacklists of sensitive words independently since they have the resource to do so, then individual developers are supposed to have a common blacklist as its length are often at tens of thousands level. To find out that, (Knockel et al., 2018) scrapped suspected blacklist files from Chinese open source projects hosted on Github and used machine-learning techniques to determine whether they are blacklists or not.

2.1 Scrapping Blacklists and Extract Their Contents

They used 21924 unique keywords which were found blacklisted by some popular apps and games, combined them with some other common blacklist related words appeared in the file name or the content like "dirty", "sensitive" and "forbid", all of which were observed from previous works, and searched them together on Github to get possible blacklist keywords files from all of the open source projects.

Then they developed a string extractor to extract lists of strings from the file. It supports most-chosen file formats for blacklists by developers such as XML, JSON and CSV, plus C-like source code and plaintexts. However it didn't support SQL database files as their grammar is not consistent.

[1] `https://freedomhouse.org/report/freedom-net/freedom-net-2017`
[2] `https://en.wikipedia.org/wiki/Censorship_by_Google`
[3] `https://www.iwf.org.uk/news/iwf-response-to-prime-ministers-statement`

2.2 Blacklists Classification

To futher classify if the list is a true Chinese blacklist, they used both anaive approach and a machine-learning approach. In the naive approach, researchers simply counts any list that contains "法轮"(Falun Gong) as a blacklist as this word appears in every blacklist publicly known. The limitation of the naive approach is also obvious since it can only find blacklists containing "法轮". Even though by far it appears in every blacklist, it can not be a guranteened feature of every possible Chinese blacklist.

In the machine-learning approach, they "used a one-class support vector machine (SVM) (Schölkopf et al., 2001), as implemented by Scikit-learn (Pedregosa et al., 2011) and LIBSVM (Chang and Lin, 2011)." The classifier requires to be trained by only one class yet maintain the ability to classify both the positive and the negative results of that class. The training data was the Chinese blacklists that were reverse engineered from popular apps, plus a blacklist from Google [4].

They modelled their possible blacklists to be "vectors of counts of the number of occurrences of each Chinese word in that list."(Knockel et al., 2018) In addition, to make the calculation tractable, they decided to apply singular-value decomposition to reduce the dimensinality of the dataset. After comparing with the golden standard, which is their result from the naive approach, the final dimension is 46.

2.3 Result

After manual verification, the naive approach returned 884 Chinese blacklists while the machine-learning approach found 1054. The longest blacklist had 38237 words and the mean length and the median length was 2128 and 1026, respectivelly.

These 1054 blacklists contains a wide range of words including erotica, Falun Gong reference, possible independence movements, criticisms to the government, political leader names, etc.

The inter-similarity between blacklists, measured by Jaccard similarity and $max(\%$ of x in y, $\%$ of y in $x)$ for list x and y, were very low. It indicates that there was very little overlap between each blacklist. It corresponds to the same discovery from the blacklists in popular Chinese applicaitons and games. Although how so many individual developers can compose blacklists on their own independently remains unknown, again, both the result from this study and the previous ones had proved that the censorship system in China is a decentralized one.

3 Characteristics of Censorable Language Beyond Blacklists

So far we have covered the first form of censorship in China, sensitive keywords blacklist. But sensitive words does not always get removed, some of them can survive. Moreover, sometimes sentences containing no blacklisted words can also be censored. In order to have strict controls over their user-generated contents, major social network services such as Sina Weibo, has incoporated manual checking in addition to blacklists to make the decision of whether to delete a post or not. So what kind of words will be removed besides those on the blacklsit already? (Ng et al., 2018) answered this question by explore the linguistic characteristics on censored and uncensored posts from Sina Weibo which contains the same sensitive words. Based on the discoveries from Psychology that rejecting information takes more efforts for human beings than accpeting it(Lewandowsky et al., 2012), and people's opinion towards are unstable overtime and contexts(XU, 2008), their hypothesis were, *a*) uncensored contents are easier to digest and *b*) the linguistic characteristics of censored and uncensored contents are different.

3.1 Corpus

The corpus consists of both censored and uncensored posts on Sina Weibo whose topic are scandles happened or happening in China. It is inspired by the Grass-Mud-Horse Lexicon, which is widely used by Chinese netizens in their sarcasms and was proved to be relavent and significant by previous studies such as (Tang and Yang, 2011) (Wang, 2012). The team chose words form the following four topics,

[4] `https://caiguanhao.wordpress.com/2012/06/01/google-gfw-blacklist`

a) pollution and food safety, *b*) internet censorship and propaganda, *c*) Bo Xilai and *d*) kindergarten abuse. All of the uncensored posts came from search results from Sina Weibo itself using the keywords from the lexicon, while censored posts were searched on FreeWeibo and WeiboScope as they track posts disappeared on Sina Weibo.

All together they have collected 1023 censored posts and 1138 uncensored ones from the four topics above, segmented by Jieba before futher experiments.

3.2 Extract Features

Kei and his team extract several features to measure the readability of a Weibo post, such as character frequency, word frequency, sentiment (only shows if the post is positive or not), sentiment classes, etc. Some of the features are unique to Chinese such as idioms, where they have also discoveried that the more idioms in a post the harder it is to understand. Finally, there are also some composite features like Readability 1 ("the mean of character frequency, word frequency and word count to semantic groups ratio", the lower the score is, the higher the readability is) (Ng et al., 2018).

To reveal more details about the linguistic characteristics, the team in addition used word embeddings as one of the feature. They computed a 200 dimensional vector for each word in a post as its representive in the large word vector space trained by 30000 latest Chinese Wikipedia aritcles. Then they did eigen decomposition for the 200*200 covarience matrix for each post. The result eigenvectors are "the directions in which the data varies the most" and the last 40 of them were used as the feature in order to reach calculation tractability since they have covered 85% of total varience.

All of these features were put into both a Naive Bayes classifier and a Support Vector Machine one to classify censored posts or not.

3.3 Result

Overall the accuracy of both the Naive Bayes classifier and the SMO classifier achieved around 0.7 in all four categories, with Readability 1 being the best performing feature. Also, the average readability score of the censored posts is lower than then uncensored ones. This suggests that readability plays an important role in censors decision as writes tend to use more uncommon words and less straight foward expressions to evade censorship. This also supports the hypothesis that uncensored contents are easier to read.

Another discovery from the experiment found that word express strong opinions, such as swear words and words that contains anger, are more likely to be censored. In other words, posts that have more potential to collective actions and social engagement are considered to be more dangerous among censors. Casual discussion on the current state of scandles survives more. By giving such an outlet to express, negative sentiment are less possible to grow into anger and leads to social movements.

4 Lanugage Encoding to Avoid Censorship

Besides studies focused on determining what kind of language are more exposed to the danger of censorship, some other studies such as the one (Ji and Knight, 2018) did tried to overview possible solutions to avoid censorship from a linguistic prospective.

For human beings encoding a language can be approached by a variety of ways. Most of them is domain-mapping, that is by mapping an entity to another entity in a different domain which is consistent and easy to remember.

Another higher level technique is story encoding. For example a post describing the struggle between the officials for the arrest of Bo Xilai goes "A few days ago, Beijing was hosting an innovative **tug-of-war** for the **elderly**; **this game** has **nine contestants** in all. **The first round** of the contest is still intense ...**The teletubby team** noticeably has the advantage and, relatively, **the Master Kang team** is obviously falling short."

All of the text in bold above refers to encoded entity names and the story it self is also a metaphor. The post also revealed three challenges for decoding. *a*) synonymy and polysemy, *b*) large number of candidates and *c*) lack of background knowledge for the target concepts and stories.

4.1 System Encoding and Decoding

Reaching the lower level language encoding techniques are already proved to be feasible to computers. (Zhang et al., 2014) and (Zhang et al., 2015) developed several approaches to automatically encode entities including Phonetic Substitution, Spelling Decomposition, Nickname Generation, Translation and Transliteration and Historical Figure Mapping. For encoding longer texts, using a simple cipher can encode and create languages like Leet and Martian script, just like encryption. Moreover, natural language generation is applied to encoding in these three ways. *a*) portmanteau neologism creation, *b*) dynamic phrasebooks and *c*) poetry passwords.

Portmanteau neologism creation (Deri and Knight, 2015) is creating new words such as mixing *friend* and *enemy* to *frenemy*. It can not be processed by machines yet and it requires a carefully picked spelling and a fusion on the phonetic level.

Next, dynamic phrasebooks (Shi et al., 2014) is something appears on the crack guide of speaking a new language for tourists. They need to know nothing of the new language but by imitating the pronounciation of the new language using phonetic spellings in another language that they already know, e.g saying *Good morning* by pronounce the Chinese words 古德莫宁 since they have similar sounds to non-native speakers.

Finally, poetry passwords (Greene et al., 2010) refers to an old technique of remembering long passwords/numbers–making a poem for it. We could imagine that by applying it in the opposite way, we could generate long passwords and use them instead of the words we are supposed to spread.

5 Summary

Censorship is a Natural Language Processing application gaining more and more focus on, especially with the rising concern about the degenerating internet freedom over the years.

People studied the linguistic characteristics of censored languages to understand the mechanism and the structure of the censorship system, which corresponds to many previous studies from other disciplines like Psychology, Computer Science.

On the other hand, Natural Language Generation tries to propose methods for people to evade such kind of censorship. It works just like the opposite of Word Sense Disambiguation–We intentially make the word sense vague and hard to decode for outsiders in order to preserve communicating any kind of information freely. People have already made some initial process on it.

Since China is the worst abuser of internet freedom there are lots of works focused on Chinese language specifically. But they all have the same limitation that there aren't too much living corpus available since the censorship system is built to remove unfavored information from day one. Also, not only the language is envolving with time, that censorship system is also improving itself at the same time. We should all bear that in mind.

References

Brew, C., Feldman, A., and Leberknight, C. (2018). Proceedings of the first workshop on natural language processing for internet freedom. In *Proceedings of the First Workshop on Natural Language Processing for Internet Freedom*.

Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):27.

Deri, A. and Knight, K. (2015). How to make a frenemy: Multitape fsts for portmanteau generation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 206–210.

Greene, E., Bodrumlu, T., and Knight, K. (2010). Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 524–533. Association for Computational Linguistics.

Ji, H. and Knight, K. (2018). Creative language encoding under censorship. In *Proceedings of the First Workshop on Natural Language Processing for Internet Freedom*, pages 23–33.

Knockel, J., Crandall, J. R., and Saia, J. (2011). Three researchers, five conjectures: An empirical analysis of tom-skype censorship and surveillance. In *FOCI*.

Knockel, J., Crete-Nishihata, M., Ng, J. Q., Senft, A., and Crandall, J. R. Every rose has its thorn: Censorship and surveillance on social video platforms in china.

Knockel, J., Crete-Nishihata, M., and Ruan, L. (2018). The effect of information controls on developers in china: An analysis of censorship in chinese open source projects. In *Proceedings of the First Workshop on Natural Language Processing for Internet Freedom*, pages 1–11.

Knockel, J., Ruan, L., and Crete-Nishihata, M. (2017). *Measuring Decentralization of Chinese Keyword Censorship via Mobile Games*. Munk School of Global Affairs.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., and Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, 13(3):106–131.

Ng, K. Y., Feldman, A., Peng, J., and Leberknight, C. (2018). Linguistic characteristics of censorable language on sinaweibo. *arXiv preprint arXiv:1807.03654*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7):1443–1471.

Shi, X., Knight, K., and Ji, H. (2014). How to speak a language without knowing it. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 278–282.

Tang, L. and Yang, P. (2011). Symbolic power and the internet: The power of a ‘horse’. *Media, Culture & Society*, 33(5):675–691.

Wang, S. S. (2012). China’s internet lexicon: Symbolic meaning and commoditization of grass mud horse in the harmonious society. *First Monday*, 17(1).

XU, J. (2008). When thinking is difficult metacognitive experiences as information. *Frontiers of Social Psychology*, page 201.

Zhang, B., Huang, H., Pan, X., Ji, H., Knight, K., Wen, Z., Sun, Y., Han, J., and Yener, B. (2014). Be appropriate and funny: Automatic entity morph encoding. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 706–711.

Zhang, B., Huang, H., Pan, X., Li, S., Lin, C.-Y., Ji, H., Knight, K., Wen, Z., Sun, Y., Han, J., et al. (2015). Context-aware entity morph decoding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 586–595.