2023-03-17

Creating a Homophone-Based Chinese Language Censorship Evasion Tool

Emily Quan

Honors Thesis

CREATING A HOMOPHONE-BASED CHINESE LANGUAGE CENSORSHIP EVASION TOOL

by
Emily Quan

Submitted to Brigham Young University in partial fulfillment
of graduation requirements for University Honors

Electrical & Computer Engineering Department
Brigham Young University
April 2023

Advisor: Dr. Philip Lundrigan

Reader: Dr. Steve Richardson

Honors Coordinator: Dr. Karl Warnick

ABSTRACT

CREATING A HOMOPHONE-BASED CHINESE LANGUAGE CENSORSHIP EVASION TOOL

Emily Quan

Electrical & Computer Engineering Department

Bachelor of Science

As the scope of Chinese language censorship expands, individuals will seek to bypass such censorship efforts. One of the most prevalent techniques in such censorship is automated keyword filtering. This research focuses on building a command-line tool that can bypass automated keyword filters for both traditional and simplified Chinese characters using a two-part approach. The first part involves detecting sensitive words in user-inputted text by using phrase matching techniques to identify character strings that have been censored in the past. The second part centers around generating possible obfuscated homonym alternatives. The tool relies on a compiled list of banned and potentially banned phrases from previous research to determine what is deemed "sensitive." Alternate characters to generate the obfuscated text are drawn from a a standardized list of the most commonly used Chinese characters. Further research is needed to automate the updating the list of sensitive phrases and to detect phrases that are similar, but not identical, to those that have been censored in the past.

Keywords: censorship, Chinese, homophone, keyword filtering

ACKNOWLEDGMENTS

To my research advisor and mentor, Dr. Lundrigan, who has been endlessly supportive; Robert Griffiths and Julie Radle, for their impactful classes that deeply influenced my college experience; Whit, my greatest champion; and, of course, to my parents, Sarah, and Marcos.

TABLE OF CONTENTS

LIST OF FIGURES

Introduction

Digital authoritarianism, "the use of information technology by authoritarian regimes to surveil, repress, and manipulate domestic and foreign populations," is gaining traction in a number of authoritarian nations (Coleman & Napolitano, 2022). Censorship is often used as a governing tool to quell dissent, prevent discussion of timely topics, and stifle opposing parties. An example of a nation-state firewall is People Republic of China's Great Firewall, a complex collection of censorship entities implemented on both the application and transport networking layers for data transmitted using the country's network (Ensafi et al., 2021). Research over the past two decades has shown that Internet censorship in China often relies on keyword filtering, or the practice of banning text that contains certain words or phrases (Rambert et al., 2021).

While non-attribution in censorship circumvention is an important technical issue often examined at varying layers in the networking stack, there is not a significant amount of research that offers solutions that enable users to share texts on the web that can evade common censorship filters (Bock et al., 2019). By developing a light-weight tool that non-technical users can deploy to modify language that may be at-risk, this solution proposes a method that would prevent users from account bans/deletions, protect users who want to discuss sensitive topics, and suggest novel ways to address timely issues at hand. The goal of this approach is to help users generate obfuscated text where the topic at hand is still clear to a stranger with knowledge of the Chinese language, and yet can bypass automated keyword filtering. This approach may not stand up to manual censorship blocks, but it would sidestep keyword filtering (Rambert et al., 2021).

Background on Chinese Language Censorship

Although digital censorship occurs in a variety of languages, it is especially prevalent in the Chinese language. Censored written Chinese words and phrases often deal with sensitive political or social topics, though there is a wide degree of variation as to what is censored (Ruan et al., 2020). The censorship takes place on a variety of platforms that cater to native Chinese language users, from gaming chatrooms to Weibo (a social media app similar to Twitter) (Clark, et al., 2017). The level of censorship and material that is censored varies widely across user location and platform, as well as the current social and political climate.

A common misconception is that Chinese language censorship is entirely performed by the Great Firewall of China (GFW). The reality is that the GFW is not a monolith, but rather a combination of different tools and tactics used to accomplish censorship (Ruan et al., 2020). Censorship is performed both through automation and through manual detection of controversial posts (individuals who read posts or entries and flag potentially problematic material) (Economy, 2018). Censorship is also not performed uniformly across various geographic regions; research shows that traffic between certain regions experiences greater censorship than in others (Rambert et al., 2021).

Certain written Chinese keywords that center around sensitive topics are perennially censored, while other keywords that have to do with newly controversial topics may not be permanently banned. Historical events that have been politically sensitive for the PRC are often perennially banned, while current trending topics that may be sensitive or controversial may only introduce bans for a finite period of time. For

example, during the Peng Shuai incident[i] in 2021, even the word 网球 (tennis) and the emoji of a tennis ball were removed at one point; however, these would not be permanently banned, since they are common words used in conversation (Zhong, 2021). Additionally, during sensitive political events within the PRC, such as the Communist Party Congress, certain keywords are banned more frequently leading up to, during, and after the event, but do not stay banned after a certain period of time (Ruan et al., 2020).

Certain communities that often deal with sensitive subject matters that are often censored are especially prone to using homophones and alternate phrases. For example, LGBTQ+ Chinese activists often experience increased surveillance and scrutiny from public officials in the digital space. Individuals who belong to these communities often use these forms of double-speak as a method of self-preservation and communication and use other methods (such as including pictures of rainbows in photo backgrounds, which are difficult to auto-detect) to continue dialogue online (Bernot, 2022).

The examination of Chinese language censorship may have primary relevance for networks located within mainland China, but its future application will extend to other countries which have implemented censorship under the guidance of the PRC, especially those that have been included in the Digital Silk Road initiative (Woodhams, 2019). Evidence suggests that some countries in Africa are following China's lead: for example, the Tanzanian deputy minister of communications has noted that although they have not yet taken the step of producing homegrown alternative sites, they intend to "guard against their misuse" (Woodhams, 2019). Censorship evasion, then, will continue to become a larger field as digital censorship increasingly becomes a global phenomenon.

The Role of Homophones in Chinese Language Censorship Evasion

Hanyu pinyin is the official romanization system for Standard Mandarin Chinese, with widespread adoption across native Chinese speakers and Chinese language learners alike, due to its usage as the primary input method for Chinese text on computers/phones. Because there are a limited number of sounds that Chinese speakers use, there are many Chinese characters that correspond to each possible pinyin romanization. This means that there are many possibilities of alternate characters, and although the tone or inflection used to pronounce the word or phrase is often different, the sounds themselves are the same.

Homophones are very common in the Chinese language and are embedded in Chinese cultural practices. For example, during the Chinese Lunar Festival, the most important holiday of the year, a number of traditions stem from puns that are symbolic of good fortune and luck. An example of this is that many Chinese people will post the character "副" on their doors upside down, because the phrase "福到了" (Fú dàole), which translates to "Fortune has arrived," uses the same sounds as "福倒了" (Fú dàole), which translates to "Fortune is upside down." During the Chinese Lunar Festival, families will often eat together and serve fish. It is a common tradition to leave some of the fish uneaten. This is because the phrase "年年有余" (Nián nián yǒuyú), which translates to "You will have surplus every year," sounds the same as "年年有鱼" (Nián nián yǒu yú), which translates to "You will have fish every year."

There are currently a number of circumvention techniques actively used by Chinese social media users. Users will include text in pictures, which is harder to detect through automated techniques. Hong Kong protestors have also been recorded as using "Hongish," or using Chinese transliterated into English to evade censorship.

The technique of using similar-sounding phrases to talk about controversial topics is already used successfully by a number of homophones popularized on Chinese social media platforms to evade censorship. For example, 河蟹 (hé xiè), which translates to "river crab," has the same phono-semantics as 和谐 (héxié), which is a censored word that refers to the policy of a "harmonious society." Some netizens have also used multimedia images (i.e. images with censored text embedded in them), but this tool will focus on altering plaintext as other forms of media are anecdotally subject to much stricter censorship guidelines.

The landscape of Chinese language censorship enforcement and evasion is often shifting, and users often develop clever ways to ensure that controversial speech continues to evolve. For example, after the #MeToo hashtag was censored in China, Chinese feminists moved towards hashtags like #MiTu (the pinyin pronunciation for the English sounds of "Me Too") and #米兔 (pronounced "mǐ tù "). This eventually evolved to #RiceBunny, which is the literal translation of the characters 米兔 (Bernot, 2022).

This is simply one example of how users cleverly circumvent filters using obfuscated keyword filtering in Chinese speech. Another example of reactionary censorship that took place in recent events (2023) involved an air balloon that floated over United States

territory. This resulted in significant interest and discussion on Chinese social media, leading to the hashtag "Wandering Balloon" to be banned on Weibo due to "relevant laws and regulations" (Wang & Dong, 2023).

Current Chinese Language Censorship Evasion Research Efforts

There are a number of ongoing research efforts focused on the technical aspects of censorship evasion. For example, University of Maryland researchers have created Geneva (Genetic Evasion), a novel experimental genetic algorithm that alters packet-manipulation-based censorship evasion strategies against censors at the nation-state level (Bock et al., 2019). Geneva works at the network level and operates by evolving its methods of network traffic alteration so that censors have a difficult time detecting its usage. Although Geneva's evolving network methodology is a compelling solution, it does not address some of the use cases that this tool addresses. Firstly, although a user's traffic would be protected down to the network layer, Geneva is limited to running on the user's traffic – meaning that if a user posted content with sensitive information, that content would still be at risk of being removed. Along this reasoning, it becomes clear that Geneva is optimized for operating on the network traffic of individuals but cannot protect a larger discussion space from being blocked.

Another censorship evasion effort is Great Fire, which is an anonymous organization based in China that seeks to make previously censored content widely accessible (Great Fire, 2018). An example of one of its services is FreeWeibo, a clone of the popular Chinese microblogging site, Weibo – only FreeWeibo contains real-time copies of over 300,000 censored and deleted posts from the site. It also produces open-source tools that make Android apps resistant to censorship and that redirect users

accessing blocked websites to mirror sites. Great Fire is also the largest aggregator of censored keywords and websites (Great Fire, 2022).

Although Great Fire's tools cover a number of services, its scope is more focused on protecting and enabling individuals than examining Additionally, since many of its tools work using redirection and using mirror websites, it utilizes redirection and alternate websites to disseminate information but does not extend to enabling continuity of conversations on the same platform. For example, a user interested in a sensitive keyword would likely be able to find the information on one of Great Fire's websites but would not have the tools to initiate a conversation about it on a censored website without it being blocked.

Research Focus

The focus of this tool's approach will be to alert a user of sensitive words or phrases that would be censored on a Chinese language platform, and then generate possible homophonic characters for the user to replace these words/phrases. While the characters will look different, they will still convey the same pronunciation of the word/phrase for a native Chinese speaker reading the obfuscated text. At the same time, however, since the alternate suggestion uses different characters, keyword filters would fail to pick up on the difference.

This tool will examine phrases from this list that are tagged as Han-simplified, Han-traditional, Han-simplified Latin (meaning that it involves both Han simplified characters and Latin characters), and Han-traditional Latin. The decision to not incorporate phrases from other scripts, such as Arabic, is due to the fact that this tool focuses on Han character replacement — taking into account phrases exclusively in other

scripts would likely require translating these phrases and suggesting an alternative word in that phrase, which is outside of the scope of this work. Due to parsing issues, this tool also does not consider phrases that combine Han characters with languages other than Latin characters. This is also due to the fact that individuals using combinations of multiple languages are often inputting search strings meant to evade censors in the first place, which is likely outside of the intended use case of this tool (in which a user is submitting text that belongs to regular sentence structures, etc.).

Because this tool is currently using a static dataset to perform its check of sensitive keywords, its word list will not include banned words that may be recent additions. The primary focus of this tool is on perennially banned topics instead of topics that may be banned because of their trending status. Since keywords based on recent events are generally more variable, it is easier to produce consistent testing results, given that keywords that may be banned during a sensitive time period may not be censored later. Additionally, since the dataset used only represents words from a 2021, it will not include keywords from events that occurred after this year (Rambert et al., 2021). However, although the primary focus of the tool is topics that are continuously censored, the tool is designed to be flexible and therefore can easily incorporate new banned phrases that may arise as well.

Although this tool will take into account both Han-simplified and Han-traditional characters, it should be noted that the vast majority, if not all, of the phrases derived from the Rambert study are phrases used by native Mandarin speakers (from both mainland China and Taiwan) (Rambert et al., 2021). Although speakers of the Chinese Cantonese dialect also use Han-traditional script, it should be noted that because of variations in

how Cantonese speakers use the script, their discussion of sensitive topics often evades censors (Hui, 2022). Therefore, this tool will be primarily useful for individuals who have been taught by Mandarin Chinese speakers.

Methodology

Determining Sensitive Keywords

To detect words that would be potentially banned within the Chinese Firewall, this tool draws upon a Carnegie Mellon University study in which researchers drew up a list of several thousand potentially banned phrases using lists from Wikipedia sourcing, Great Fire keyword lists, and the knowledge of native Chinese speakers and sent these phrases between servers located in different geographic areas both inside and outside China to see which phrases were banned/blocked (Rambert et al., 2021). The study characterizes these phrases by the script that they use, such as Arabic, Latin, Han-simplified or Han-traditional, etc. Some of the phrases involve a mixture of several different scripts and are characterized accordingly.

Several other datasets were considered for examining banned phrases. For example, the China Fire project has compiled an extensive list of phrases that Internet users have noticed are banned. The list also includes Wikipedia articles that were previously banned in China — Wikipedia has been wholly banned in China since 2019, but prior to that, only select articles were banned (Harrison, 2019). However, the Rambert study phrase list was chosen because it also takes into account many of the entries from the China Fire project and includes some novel phrases that were tested as potentially sensitive (Rambert et al., 2021).

Tool Implementation

This tool was written in Python with the assistance of pinyin, a Chinese language library, pandas, a popular data analysis library, and spaCy, an industrial strength Natural Language Processing (NLP) library. The tool was implemented as a command-line

interface (CLI) because it allowed the user to customize inputs and settings as necessary while also creating an easy package for users to transfer and use. While a browser extension may be more familiar to some users, it would also be easier for a host country looking to ban such sensitive applications to restrict its usage.

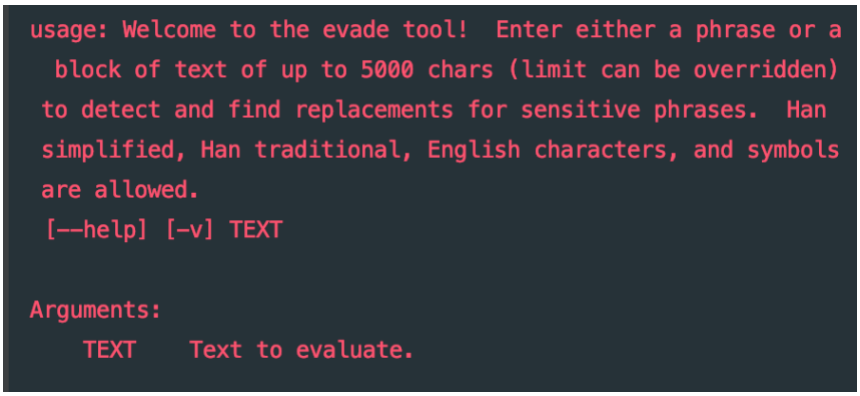

Figure 1. A screen capture of the usage information for this tool.

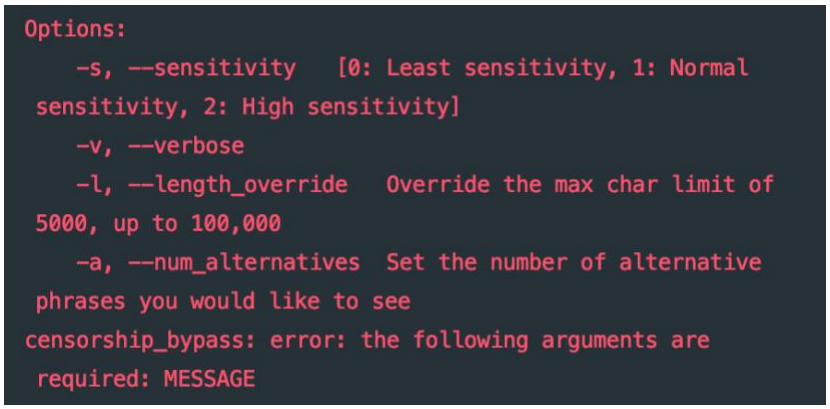

Figure 2. A screen capture of the tool options presented to the user.

The tool first takes in the keyword phrases from the Rambeau study and cleans the data as needed. Several of the entries from the dataset contained anomalous entries, some of which seemed to be test entries (some of the words that were tested were simple words, like 山 · which means mountain). For simplification purposes, these entries were deleted. The tool then also only takes into account phrases that fall into the script categories that are considered for this study (variations of Han-s and Han-t with possible

Latin character additions). Some of the phrases that fall within the scope of this tool additionally have special characters, such as + signs (in order to catch search requests that would be censored). Although it is unlikely that the target users of this tool would input phrases that include special characters, phrases in this category are also included in the process. In order to avoid overly long processing times, there is a default 5000-character limit imposed on user input. This can be overridden to 10,000 characters by the user. Also, since many of the phrases from the Rambeau dataset are several characters long, the tool uses the spaCy library to perform phrase matching in order to optimize performance.

The tool can be run with different sensitivity levels. The Rambeau study tested each phrase between 92 different city links that involve cities both inside and outside of mainland China. The results of each phrase (pass, blocked_update, always_blocked, blocked_search) are recorded in the dataset. Running the tool with a low sensitivity level will only cause the tool to search against phrases that triggered results in the "always_blocked" category. A normal sensitivity level will search against phrases that produced any type of blocking. A high sensitivity level will search against any of the phrases included in the dataset, including ones that did not result in any detectable censorship.

Additionally, running the tool with the verbose mode allows the user greater insight into possible alternatives. When the verbose mode is run, users can see all of the possible pinyin equivalents for each of the characters that will be modified, as well as each alternate character's popularity ranking (how often it is generally used in written Chinese).

Identifying Alternatives

In order to find character substitutes, a word list of the most common 9,900 characters used in the Hanyu Shuiping Kaoshi (HSK) Chinese test was ingested so that substitutes could be drawn from this list. The HSK test was used, since it is a very common standardized test for non-native speakers of Chinese, and its list of common Chinese characters is frequently used. One alternative implementation considered was to use Google Translate HTTPS requests to derive different variations of the text. However, this approach would have a larger time cost in comparison to a static dictionary lookup. Furthermore, the HSK word list includes the pinyin for each character entry with tones, meaning that characters can be selected based on the tonal value as well. This translates to greater understandability for users when they are reading the subsequent obfuscated phrases.

When a sensitive phrase is detected in the user's input text, the tool parses the phrase character by character to determine alternates. Since most of the keyword phrases tested involved more than one character, the homonym replacement only takes place for every other character in the string. For example, if the keyword phrase was "北京 政变" [4 characters long], the tool will suggest replacement phrases that alter the second and fourth characters [京 and 变] but leave the first and third characters [北 and 政] intact. This is to make longer phrases still relatively readable for users, while ensuring that an automated keyword filter will not pick up on the exact phrase. Additionally, since many Chinese names are three characters long, replacing every other character starting with the second character will only replace the middle character, leaving a well-recognizable name easy to identify.

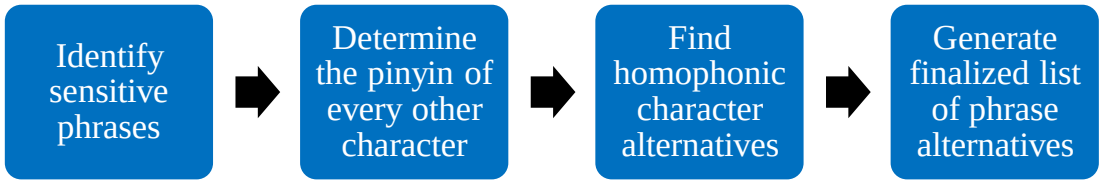

Figure 3. The process by which the text is scanned and alternate phrases are generated.

For each character that will be replaced, the tool determines the pinyin value for a character that will be replaced, then searches the HSK list to find characters that match this value. The pinyin for the sensitive phrase is determined using the Python pinyin library, which generates the pinyin value for the phrase. The tool defaults to generating three alternatives (provided there are three alternative characters possible for each variation), but this number can be increased by the user. In addition to the alternates that involve finding Han character homonyms, the tool also produces alternates that intersperse pinyin and Han characters, resulting in the use of "Hongish" type phrases.

Testing

Testing was performed for specific use cases, such as whether the program correctly produces homonyms for single characters, input with special characters, input in multiple languages, and multi-character phrases. The use cases are listed in the table below, and the specific phrases used for each can be found in Table 1.

The general criterion for this testing is simple: does the tool correct suggest an alternative that acts as a homonym, does it correctly perform exception handling, and does it include pinyin alternatives for the user?

Although it may seem logical to perform tests on state infrastructure that performs the censorship, this category of testing will be outside of the scope of this thesis for several reasons. The first is that such testing in previous studies (such as the Rambeau study) often involves the usage of infrastructure physically located in specific cities, and such infrastructure can be difficult and costly to procure. The second is to mitigate the inherent risk involved in taking action that could be perceived as adversarial against other governments or states. The third is that the purpose of testing performed on this tool is mostly to verify the results of that the output is generated as desired. The readability and ease of use of alternatives from this tool is reserved for future study and research.

Sample Results

```
Sensitive phrase detected: Keyword_List
span.text: 骚
pinyin of span.text: sāo
pinyin detected the same:       frequency_rank character  ... hsk_level general_standard_num
2336            2237         骚  ...       6.0              2823.0
3553            3454         搔  ...       NaN              2539.0
4488            4389         臊  ...       NaN              3423.0
4646            4547         缫  ...       NaN              5848.0
8007            7908         鳋  ...       NaN                 NaN
8829            8730         鼜  ...       NaN                 NaN
9013            8914         鳋  ...       NaN                 NaN

[7 rows x 9 columns]
Alternate phrases:
character    搔
Name: 3553, dtype: object
character    臊
Name: 4488, dtype: object
character    缫
Name: 4646, dtype: object
```

Figure 4. Output for a single character replacement.

Figure 4 shows the output for a single character replacement, shown here in verbose mode. The tool detects the pinyin of the character, then identifies HSK replacements based on the tone and phonetic value. It then selects the top three alternate values that a user could utilize in this instance.

Another example is for homonyms produced for the phrase “北京政变” (Běijīng zhèngbiàn). Based on the principle of exchanging every other character, the tool generates the following results: “北经政便” (Běi jīng zhèng biàn) and“北精政遍” (Běi jīng zhèng biàn). Although the meaning is entirely different for an automated keyword filtering tool, the reading of the characters sounds the same as the original.

```
Sensitive phrase detected: 北京政变
pinyin of phrase: běijīngzhèngbiàn
Phrases after removing duplicates: ['北经政变', '北京政遍', '北精政便']
Sensitive phrase detected: 政变
pinyin of phrase: zhèngbiàn
Phrases after removing duplicates: ['政便', '政遍', '政辩']
```

Figure 5. Output generated when a sensitive phrase is entered into the tool.

Other Considerations

There are several instances in which there are sensitive phrases located within other sensitive phrases. For example, the phrase "民远" (Mín yuǎn) is listed as a sensitive phrase, but so is the entry "民远 1989," which contains the previous phrase. In these instances, the tool detects both phrases but will use the value of the shortest phrase to determine alternatives. This maintains maximal integrity of the text while still offering a replacement for the sensitive keyword so that it can be obscured.

In rare instances in which there were no character alternatives available for the specific phonetic-tone combination, the default character value was used. Since data analysis an estimated 1345 pinyin combinations used in written Chinese, it is generally rare for a character to not have a common alternative and therefore was determined that trying to replace the character with a similar-sounding (but not exact) substitute would be more of a hindrance than an aid (Chinese Pinyin Combinations). Conversations with a Chinese language specialist ultimately resulted in the conclusion that since tone marks are integral to differentiating various characters, it was decided for the first iteration of this tool to incorporate less alternatives that maintain the fidelity of original tone markings instead of generating more alternatives that only maintain the phonetic sound

(Guo, 2022). Additionally, since it is rare that individual characters are censored (since many characters are used in multiple contexts or compound words or phrases), this fringe case is not generally an issue for users looking to implement censorship circumvention.

Some characters in written Chinese have multiple pronunciations under different contexts. For example, the character 了 has several possible pronunciations, depending on the context: le (often used to denote past tense), liǎo (to finish or conclude), and liào (to watch from a height or distance) (Pleco). The Python library used to determine pinyin for this tool is the Python pinyin library, which uses a basic lookup strategy using a Mandarin.dat file derived from the CC-CE-DICT project, a public domain Chinese-English dictionary (mdbg, 2020). The dictionary includes multiple pinyin options for characters with more than one pronunciation, but only the first (most common) pronunciation is returned on API calls.

Figure 6. Illustrating how the Python pinyin library does not select the character pronunciation on basis of the context of the character.

Choosing which pronunciation to return is a common difficulty encountered in machine translation; note that the phrase 不得了, which contains the character 了, should return the pinyin value Bùdé liǎo when entered in Google Translate, but instead returns the incorrect pinyin value, Bùdé le (using the most common pinyin value instead of the correct value) (Google Translate, 2023). Since this accounts for greater inaccuracy in the tool (since characters will not necessarily be replaced with their appropriate

pronunciation in context), determining the pronunciation of a character in context is an area of future research.

Since this tool will be accessible to the general public, individuals could theoretically reverse engineer the tool to add all of the generated alternatives of banned keywords to existing censorship platforms, nullifying the usefulness of the tool. However, determining an ideal keyword substitute is beyond the scope of this particular problem. The purpose of his tool is not meant to select the best alternative – it merely generates possible suggestions to empower users to determine what makes the most sense in various contexts. Since users have the opportunity to generate different numbers of alternatives, they can also modify generated alternatives to fit their uses. Additionally, governments must strike a delicate balance with censorship tools, since the goal of censorship is not to create a sterile online environment free of controversial topics, but rather to discourage the community-building of coalitions that may become problematic (Zhong, 2021). As a result, as long as there remains online discourse in written Chinese, there will remain opportunities to use alternatives in the future.

Leaving the ultimate decision of choosing alternatives to the user also makes sense since different selections make sense for in different contexts. Some users may be looking for a clever turn of phrase, while others may be simply trying to select characters that look similar on a character radical basis. Written Chinese characters are comprised of radicals, which are components that may have phonetic relations. For example, characters 们 (men) and 们 (men) both have the same phonetic pronunciation due to a radical they share. Users trying to obscure texts for an automated filter but who want to

keep their texts relatively easy to read may be more interested in choosing similar examples that have similar radicals.

Additionally, presenting alternatives to the user without automatically replacing the characters also provides an opportunity for the user to consider the inherent risks of taking a homophonic approach. In 2022, Weibo administrators announced that they were increasing efforts to clean up "misspelled" words, homophone characters, and variants of words (Koetse, 2022). Content that falls under this category under Weibo's rules may be deleted, and this may cause a user's account to fall under increased scrutiny or to experience restrictions. Because of this, users should carefully consider how they choose to use this tool and the implications that their actions may have.

Discussion

Future Research Areas

There are numerous areas for additional research in Chinese language censorship evasion. The first obvious inclusion would be adding additional support for different languages that currently experience keyword filtering, such as Arabic, Russian, and Japanese. Additionally, although users have not reported current censorship restrictions for Cantonese speakers, expansion of evasion in this area may be helpful as censorship rules evolve. This would generally require using additional libraries that cater to parsing various languages, including using a library that would be specific to Cantonese (a common library used for this is PyCantonese).

Automating the addition of future sensitive keywords to check for that arise based on political or social events would also be extremely helpful, since the current model requires manual updates to the list. This could be done through the automation of detecting trending topics that are controversial on social media sites such as Twitter or Weibo. A controversy score could be determined by periodically testing whether popular phrases later experience restrictions while being sent. Since the majority of blocked phrases experienced restrictions when they were sent between domestic mainland Chinese cities, this would likely require gaining access to and maintaining Chinese infrastructure, which would require extended effort (Rambert et al., 2021).

There are also future areas for testing this tool in real-world applications. For example, it would be interesting to examine how long a modified homophone that expresses a controversial topic would stay up on Chinese social media before getting banned. Various generated homophone alternatives could be scored based on both how

effectively they evade censorship and the degree to which they are re-used/spread by users. Performing a user test with the generated homophones with both native and non-native Chinese speakers to determine the intelligibility of phrases would also be informative, since both groups would have different experiences with pinyin and reading characters. Additionally, it would be instructive to study the degree to which incorporating homophones hampers one's ability to read a text.

In light of increased censorship and scrutiny on Chinese social media sites, users have also resorted to using words and terms that are not exact homophonic matches but instead have some relation to the original word's meaning. For example, in order to discuss recent bank protests in Henan, some individuals started using the word for Netherlands in Chinese (which is pronounced Hélán) as a way to mention the events in Henan (Koetse, 2022). This led to other individuals using code words that built on the idea of using Holland as a replacement for Henan. For example, a nearby Chinese city, Zhengzhou, was referred to as "Amsterdam" in the discourse, and bank deposits were described as "tulips" (Koetse, 2022).

A future iteration of this tool could build on this idea by selecting a main keyword to modify and then generating suggestions of related words to utilize, or by ensuring that keyword alternatives maintain the same parts of speech. For example, a user looking to change a date or a location would be presented with an alternative date or location as a replacement, even if the pinyin is not an exact 1-to-1 exchange.

In terms of improving the tool's classification model, a further update would likely include finding close (but not exact) matches of various phrases using a tokenization model. This would help users ensure that phrases that were not included in

the original keyword list or that may be censored in the future would continue to evade current censorship efforts. The tool would also greatly benefit from a model which would try to select alternatives that would preserve meaning as closely as possible while still preserving pronunciation (and still maintain the capability to bypass automated censoring).

Conclusion

As the scope of digital authoritarianism spreads across country borders, censorship evasion will become more relevant for an increasing number of individuals across the globe (Coleman & Napolitano, 2022). Capitalizing on unique features in various languages introduces opportunities for novel approaches to bypass state-established restrictions and enable the discussion of sensitive topics. This tool provides the beginnings of a technical solution to evading automatic Chinese language keyword filtering, and future iterations can improve its complexity and scope to match current censorship evasion needs. As sophistication in both enforcement and evasion continues to evolve, there are important safety considerations for users to consider in deciding whether to use homophonic alternatives. However, in an age where the Internet remains a global space for individuals to connect and learn, censorship evasion can ensure that future generations remain informed.

Bibliography

Bernot, A. (2022, November 20). *Double-speak as Resistance to LGBTQI+ Repression in China*. Retrieved from The China Story: thechinastory.org/double-speak-as-resistance-to-lgbtqi-repression-in-china/?ref=china-neican

Chinese Pinyin Combinations. (n.d.). Retrieved from Drive.google.com/file/d/1UrwkxWZMISUDy493deipSxzJsqDCtiik/view

Clark, J., Faris, R., Morrison-Westphal, R., Noman, H., Tilton, C., & Zittrain, J. (2017). The Shifting Landscape of Global Internet Censorship. *Berkman Klein Center for Internet & Society Research Publication*.

Coleman, V., & Napolitano, J. (2022, March 14). Digital Human Rights Need a Single Home in U.S. Government. *Foreign Policy*.

Economy, E. C. (2018, June 29). The great firewall of China: Xi Jinping's internet shutdown. *The Guardian*.

Geneva. (2022). *Geneva: Evolving Censorship Evasion*. Retrieved from censorship.ai: geneva.cs.umd.edu/

Google Translate. (2023, February 1). Retrieved from Google Translate: translate.google.com/?sl=auto&tl=en&text=%E4%B8%8D%E5%BE%97%E4%BA%86&op=translate

Great Fire. (2018). Retrieved from Great Fire: en.greatfire.org

Great Fire. (2022). *Censorship of Weibo Searches in China*. Retrieved from Great Fire: en.greatfire.org/search/weibo-searches

Guo, L. (2022, November 17). Importance of Tone Markings in Pinyin. (E. Quan, Interviewer)

Harrison, S. (2019, May 21). Why China Blocked Wikipedia in All Languages. *Slate*.

Hui, M. (2022, September 5). China's internet censors have a blindspot: Cantonese. *Quartz*.

Koetse, M. (2022, July 13). *Weibo Vows to Crack Down on Homophones and 'Misspelled' Words to "Stop Spread of Harmful Information"*. Retrieved from What's on Weibo: https://www.whatsonweibo.com/weibo-vows-to-crack-down-on-homophones-and-misspelled-words-to-stop-spread-of-misinformation/

mdbg. (2020, August 18). *CC-CEDICT Home*. Retrieved from CC-CEDICT Wiki: cc-cedict.org/wiki/

Pleco. (n.d.). Dictionary Entry for 了.

Ruan, L., Crete-Nishihata, M., Knockel, J., Xiong, R., & Dalek, J. (2020). The Intermingling of State and Private Companies: Analysing Censorship of the 19th National Communist Party Congress on WeChat. *The China Quarterly*, 497-526.

Wang, V., & Dong, J. (2023, February 7). China Tries to Play Down Balloon Dispute With Censorship and Memes. *The New York Times*.

Woodhams, S. (2019, February 23). How China Exports Repression to Africa. *The Diplomat*.

Zhong, R. (2021, December 5). *Peng Shuai and the Real Goal of Chinese Censorship*. Retrieved from Wired: https://www.wired.com/story/peng-shuai-censorship-china/

Table 1

Testing Various Categorizes of Entries Involving Han-s/Han-s and Latin Characters

Character	Pinyin	Alternate 1	Alternate 2	Alternate 3
骚	Sāo	搔	臊	缫
锞	kuǎ	垮	侉	銙
2000 东突入大陆	2000dōngtūrùdàlù	2000 氡突洳大录	2000 东突入大路	2000 咚突缛大露
97 吻	97wěn	97 稳	97 紊	97 刎
传真 1 -800 -608 -3158	chuánzhēn 1 -800 -608 -3158	传针 1 -800 -608 -3158	传侦 1 -800 -608 -3158	传珍 1 -800 -608 -3158
64 民主	64 mínzhǔ	64 岷主	64 缗主	64 珉主
89 的六月肆号	89delìuyuèsìhào	89 的六阅肆浩	89 地六跃肆皓	89 地六越肆耗
平*反	píng*fǎn	平*返	n/a	n/a
刘晓波+死	líuxiǎobō+sǐ	刘小波+死	刘筱波+死	刘篠波+死
淫{5}秽		n/a	n/a	n/a
习近平+川普进行了+含蓄的批评	xíjìnpíng+chuānpǔjìnxíngle+hánxùdepīpíng	习近平+氚普近行饹+寒蓄的批凭	习禁平+巛普禁行饹+函蓄地批瓶	习进平+穿普进行了+含蓄的批平
六四万岁	Lìusìwànsùi	六肆万岁	六饲万岁	六寺万岁

Character	Pinyin	Alternate 1	Alternate 2	Alternate 3
郑海天	zhènghǎitiān	郑酰天	郑胲天	郑醢天
对平民与学生使用了武力	dùipíngmínyǔxuéshēngshǐyònglewǔlì	对凭民予学升使用了午力	对平民与学生使用了五力	对评民语学声使用了武力
操纵海外特务	Cāozònghǎiwàitèwù	操縱海外特误	操糍海外特悟	操纵海外特物
木樨地 纪念	mùxīdì jìniàn	木吸地 技念	木西地 计念	木息地 记念
温彻斯特 1000x	wēnchèsītè1000x	温撤斯忒 1000x	温澈斯忐 1000x	温掣斯慝 1000x
无线窃听器 QQ	WúxiànqiètīngqìQQ	无线窃厅器 QQ	无限窃汀器 QQ	无现窃听器 QQ
chai 玲	chai líng	chai 灵	chai 龄	chai 凌
猎枪仿真枪 QQ	LièqiāngfǎngzhēnqiāngQQ	猎呛仿侦枪 QQ	猎跄仿珍枪 QQ	猎腔仿针枪 QQ
ATOM 弹制造	ATOMdànzhìzào	ATOM 旦制灶	ATOM 弹制躁	ATOM 但制造