Creative Language Encoding under Censorship

Heng Ji
Rensselaer Polytechnic Institute
jih@rpi.edu

Kevin Knight
University of Southern California
knight@isi.edu

Abstract

People often create obfuscated language for online communication to avoid Internet censorship, share sensitive information, express strong sentiment or emotion, plan for secret actions, trade illegal products, or simply hold interesting conversations. In this position paper we systematically categorize human-created obfuscated language on various levels, investigate their basic mechanisms, give an overview on automated techniques needed to simulate human encoding. These encoders have potential to frustrate and evade, co-evolve with dynamic human or automated decoders, and produce interesting and adoptable code words. We also summarize remaining challenges for future research on the interaction between Natural Language Processing (NLP) and encryption, and leveraging NLP techniques for encoding and decoding.

1 Introduction

According to a recent study by Freedom on the Net[1], 67% of netizens live in countries where criticism of the government, military, or ruling family is subject to censorship. For example, censorship on Chinese social media is intensive (Chen et al., 2013; Zhu et al., 2013; King et al., 2014; Hiruncharoenvate et al., 2015). Users must register with their real identities (name, social security number, address, phone number and email). Social media platforms temporarily shut down commenting functions or remove user accounts. One part of the censorship is conducted via maintaining and updating a list of blacklisted keywords [2] to block certain known code words. This aggressive censorship mechanism also produces false alarms and filters regular posts. For example, on July 22, 2012, a heavy rainstorm in Beijing killed 77 people partially due to the slow response from the government. General terms such as 事故 *(accident)* were added into the blacklisted word lists, and thus all posts about any accidents were removed. As another example, on April 20, 2018, a popular humorous video app *Neihan Duanzi* with more than 100 million users was permanently shut down. Angered by the closure, many protests organized online as flash-mobs, using secret signals to allow *Duanzi friends* to recognize each other: a car-horn beeped in a specific long-short-short rhythm *Di...di di*, a double-flash of the headlights, and a coded song sung at the same time across the country by replacing 孤立墙 *(isolated wall)* with 柏林墙 *(Berlin wall)*, 艳阳雪 *(sunshine and snow)* with 六月雪 *(snow fell in June, an idiom that indicates injustice)*.

In this paper we overview how humans and machines encode intent into appropriate, interesting, and adoptable language to enable netizens to evade censorship and freely communicate with general netizens. These techniques attack a fundamentally challenging problem in automatically understanding fast-evolving social media language, especially for netizens without culture-specific knowledge. Coded words have made human language evolve faster than ever, on a daily basis, due to the tension between encoding and decoding objectives. This opens an unexplored area of coded language processing.

[1] https://freedomhouse.org/report/freedom-net/freedom-net-2016
[2] https://en.wikipedia.org/wiki/Internet_censorship_in_China

Proceedings of the First Workshop on Natural Language Processing for Internet Freedom, pages 23–33
Santa Fe, New Mexico, USA, August 20, 2018.

2 Human Encoding

2.1 Categories of Coded Language

Netizens create and use obfuscated language for a variety of purposes.

Discussing sensitive information and evading censorship. Code words widely exist in Chinese social media (Huang et al., 2013; Zhang et al., 2014; Zhang et al., 2015). Bamman et al. (2012) automatically discover politically sensitive terms from Chinese tweets based on message deletion analysis. When Chinese netizens talk about the former politician 周永康 *(Zhou Yongkang)*, they use a coded word 康师傅 *(Master Kang)*, a brand of instant noodles whose Chinese spellings share one character 康 *(kang)*. The Enron emails[3] also include many code words, such as *dinosaur* referring to an illegal stock.

Masking illegal activity. In the dark web for human trafficking, arms dealing, and drug dealing, *research chemical* or *RC* is euphemistically used to discuss psychoactive chemicals that are not yet scheduled as narcotics. *ice* and *skiing* refer to *Cocaine*. Phone numbers are often obfuscated (e.g. *7o2 7two7 7four87*, *four-oh-eight 9OO-one*) (Szekely et al., 2015).

Terror plots. Terrorists often use seemingly innocent conversations laced with coded messages and double-speak. 9/11 attacker Abu Abdul Rahman told accomplice Ramzi Binalshibh in an Internet chat room: *To German girlfriend: The first semester commences in three weeks. Two high schools and two universities... This summer will surely be hot ... 19 certificates for private education and four exams. Regards to the professor. Goodbye.*[4] Here, *in three weeks* refers to *September 11, 2011*, *two high schools* refers to *World Trade Center*, *the professor* refers to *Bin Laden*, *19 certificates* means *19 hijackers* and *four exams* means *four planes*, etc.. In addition, three targets bear a code name. The US Capitol building was called *The Faculty of Law*; the Pentagon became *The Faculty of Fine Arts*; and the North Tower of the World Trade Center was code-named as *The Faculty of Town Planning*. Table 1 lists more examples of seemingly innocuous terms which are terror code words and their meanings, based on reports from FBI and the International Center for Political Violence and Terrorism Research in Singapore[5][6][7].

Code Word	**Meaning**
peanut butter, wedding, jelly sandwich	terrorist attacks
culinary school	training camps
rap concert	a training run before an attack
rash, skin disease	under surveillance
eggplant	rocket-propelled grenade
tourism, smelling fresh air, warehouse	taking part in jihad
little girl	terrorist
Abid Naseer's girlfriend Nadia	bomb
banana, creps (Nike trainers)	firearm
cheese	money
iron	weapons
football, soccer	fighting
get married	to be killed
Brian, Bob	FBI agents
land of the two rivers	Iraq
P-town	Pakistan

Table 1: Examples of Terror Code Words

[3] https://en.wikipedia.org/wiki/Enron_Corpus

[4] https://en.wikipedia.org/wiki/Ramzi_bin_al-Shibh

[5] http://insider.foxnews.com/2014/09/20/inside-language-terrorism-isis-code-words-include-pbj-rap-concert-and-rash

[6] https://www.reuters.com/article/us-usa-padilla/u-s-terrorism-trial-ponders-meaning-of-eggplant-idUSN0129968420070701

[7] https://nypost.com/2014/09/13/jihadi-tapes-reveal-sinister-pbj-code-in-culinary-school/

Plain slang in a particular domain, community, youth language, or cutesy language: In sports gambling, *getting down* means *placing a bet*. Japanese cartoon and game fans create many domain-specific code words: e.g., 電波 *(Radio wave)* refers to weird but charming characters who mind-control others by doing unexpected things. In teens' secret language, acronyms are widely used: *ABFL* means *a big fat lady*, and *POTS* means *parents over the shoulder*. Some code words are not used to conceal, but rather as trendy jargon, e.g., *ELI5* means *Explain Like I'm 5*.

Expressing strong emotion and sentiment, or making fun: Sometimes netizens encode fun and creative language to express sarcasm, irony statement, positive/negative sentiment, strong emotion, or vivid descriptions of entities and events. 剁手党 *(hand-chopping party)* refers to a group of people who are addicted to online shopping. The majority regret their shopping sprees and vow in vain to chop off their hands should they spend their money so easily again.

2.2 Challenges for Decoding

These kinds of coded language present challenges for censorship:

Synonymy and Polysemy. Mature information extraction techniques can extract and resolve entities and events from text. However, coded language defeats most contextual features used by these techniques. Less-mature metaphor detection (Wang et al., 2006; Tsvetkov, 2013; Heintz et al., 2013) techniques also have trouble with abstract coded language. Word-sense disambiguation techniques (Yarowsky, 1995; Mihalcea, 2007; Navigli, 2009) deal in sense inventories that are fairly static over time, whereas code language evolves rapidly.

Large number of candidates. Coded language effectively hides in plain sight because anything can be a candidate to the decoder. It is challenging to discover code words from social media text because less than 3% unique terms are code words (Zhang et al., 2015). Moreover, common words and phrases are often used as code words to avoid censorship. For example, 姐夫 *(jie fu)* can be used to refer to both its literal meaning *brother-in-law* or the Russian politician Medvedev whose name's Chinese transliteration ends with syllables that have similar pronunciations as *jie fu*. Likewise it is difficult to resolve to real targets after they are identified. Entity Linking techniques such as (Pan et al., 2015) resolve entity ambiguity by linking mentions to an external knowledge base (KB), while there is no existing KB that covers a wide range of target entities, implicit concepts and events referred by code words.

Lack of labeled data. No sufficient mention-level code annotations exist for training an end-to-end encoder or decoder. Manual code annotations require native speakers who have cultural background (Zhang et al., 2014), so novel approaches are needed to save annotation cost.

Lack of background knowledge for the target concepts and stories. 薄熙来 *(Bo Xilai)* , a former Chinese politician, was found guilty of corruption, stripped of all his assets, and sentenced to life imprisonment. In Chinese social media Bo is coded as 平西王 *(Conquer West King)* who was a historical figure four hundreds years ago who governed the same region as Bo and also suffered from a downfall and an arrest at the end of his career. 真红女王 *(Truly Red Queen)* is a main character in Rozen Maiden, a Japanese manga series. It is used as a code name for 江青 *(Jiang Qing)*, a major political figure during the Cultural Revolution by Chinese Communist Party; because the color red is associated with communism, and she was the fourth wife of Mao Zedong, the first Chairman of China (*queen*).

2.3 Entity Encoding

Zhang et al. (2014) did a preliminary study on categorizing methods to create code words that refer to entities. The surface forms of human created code words usually appear quite different from their target names. 26% of human encoded entity mentions are based on the entity's reputation and public perception, and 21% of them are based on modeling the entity's characteristics. According to (Zhang et al., 2013), in 1,500 randomly sampled tweets that include 10,098 unique terms, there are 1,342 coded mentions that refer to 250 unique entities. Some frequent code word examples from Reddit[8] and Sina Weibo[9], a

[8]https://www.reddit.com/

[9]https://weibo.com

China-based microblogging platform, are presented in Table 2.

Code Word	**Target Entity**	**Comment**
苏牙 (Su-tooth)	Luis Suarez	Luis Suarez bites other players.
康敏苏小姐 (Miss Coming Soon)	Lady Gaga	Lady Gaga said that her "Do What You Want" MV was coming soon several times around five years ago but it has not been released yet.
邓黑猫 (Deng Black Cat)	邓小平 (Deng Xiaoping)	*Black Cat* originates from Deng Xiaoping's famous quote: "No matter white or black, a cat that catches mice is a good cat".
Indian Hay	cannabis	from India
Jelly Bean	amphetamine	similar shape
Jelly Green	marijuana	similar color
Ice, Lady Snow	cocaine	similar appearance

Table 2: Examples of Attribute-based Code Words

Many human-created code words are based on domain mapping, done in a consistent and memorable way. Table 3 and Table 4 present some examples. From these examples we can see that the most popular code domains are food, celebrity, animal, and cartoon. These code words tend to be innocent, cute, vivid, impressive, and easy to remember. The incongruity theory (Mulder and Nijholt, 2002) states that humor is perceived at the moment of realization of incongruity between a concept involved in a certain situation and the real objects thought to be in some relation to the concept. Code words created by domain mapping naturally match this theory because of the domain shifts between code words and the contexts. For example, *kiwi* is the most popular code word referring to Chris Evans, due to his bearded face. So a post like "kiwi took off his shoes!" is funnier than "Chris took off his shoes!". Many code words are created based on archetypes, i.e., innate, universal prototypes. When such prototypes can be observed among celebrities, their names will be directly used as code words. For example, 楚霸王 *(Xiangyu, king of western Chu)* is often used to refer to brave people. When no proper names could be easily found to represent a prototype, a new code name will be created. For example, 朝阳群众 *(The people of Chaoyang district)* refers to anonymous volunteer security patrols in Beijing. They began to gain fame on social media after being credited by police as leading to the arrests of several celebrities involved in drug-related crimes.

When using the domain-mapping strategy, code words in the same thread tend to be consistently derived from the same domain, and thus they are topically coherent. For instance, in a thread when *ice* is used to refer to *methylamphetamine*, then ice or snow related code words will be used for other drugs (e.g., *Lady Snow* for *cocaine*) and actions (*ice cream habit* for *occasional use of drugs*). In contrast, when *juice* is used to refer to drugs, then *Pepsi habit* will be used instead. Similarly when animals are used to refer to drugs, then *chicken scratching* and *henpicking* are used to describe a crack addict searching on their hands and knees for drugs; and *bulls* instead of the more common *Big John* will be used to refer to the police.

2.4 Story Encoding

Obfuscated language is also used to encode an entire paragraph, especially for planning certain activities. For example, North Korean people changed the well-known rhythm song "*Three Bears*" to include satirical references to the Kim family dynasty:

- **Original version:** *Three bears in a house, papa bear, mama bear and baby bear. Papa Bear is fat, Mama Bear is thin, Baby Bear is too cute.*

Code	Code Domain	Target	Shared Characteristics
lemon, lemonade	food	poor quality drugs	light, ineffective
lettuce	innocent items	cash money	cheap items
wild cat	animal	methcathinone	methcathinone's spelling includes *cat*, strong effect (*wild*).
Jane, Jay, Juanita, Aunt Mary	common person name	marijuana	rhyming slang
King	common, person name	cocaine	rhyming slang
Jimmy	celebrity	the foil used when smoking heroin	*foil* is a rhyming slang for *Boyle* in Jimmy Boyle
Kate Bush	celebrity	Kind Bud	*Kind* is pronounced *Kine* in its origin Hawaii, which means *excellent*
Jefferson Airplane	celebrity	used match cut in half to hold a partially smoked marijuana cigarette	Jefferson Airplane was split into the two bands
LBJ	celebrity	LSD, PCP and heroin	LBJ is best known as the initials of former US president Lyndon, Baines and Johnson
420	song	marijuana	420 = 12*35, Bob Dylan's song "Rainy Day Woman #12 and 35" has a line "Everybody Must Get Stoned" which is mistakenly associated with smoking marijuana
Snow White	cartoon	cocaine	similar appearance with snow flake
Peter Pan	cartoon	PCP	similar abbreviations
Tina	TV series	crystal methamphetamine	Molly Meade is a character in "Ugly Betty", *Molly Meade* is rhyming slang for *methamphetamine* and *Tina* is created to represent as Molly's ugly and meaner (*crystal*) sister.

Table 3: Domain Mapping Examples for Drug Dealing

- **Modified version:** Three bears in a house, **pocketing everything**; **grandpa** bear, papa bear and baby bear. **Grandpa** Bear is **fat**, Papa Bear is fat, too, and Baby Bear is a **doofus**.

Chinese netizens sometimes need to encode a whole story in order to express politically-sensitive ideas. For example, a thread about the arrest of Bo Xilai was initiated by a coded post and then replied to by another post using code words from the same food domain (nine contestants refer to the nine members of the CCP Standing Committee):

- **Initial Post:** *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.*
- **Reply Post:** ***Tomato** has retreated; what **flavour** will **Master Kang** still have?*

The idea of story encoding is not new. Many classic English nursery rhymes were started as 19th-century "blog posts" criticizing British government figures. It can be further traced back to the tradition of writing acrostic poems and articles in ancient China.

Code	Code Domain	Target	Target Domain	Shared Characteristics
西红柿 (tomato)	vegetable	Chongqing City	politics	tomato is a homonym for 西红市 *(western red city)*; Chongqing, the largest city in China's southwest, was the front line of the red revival.
Black Mamba	animal	Kobe Bryant	sports	venomous, agile, aggressive.
silver fox	animal	Anderson Cooper	celebrity	gray hair, handsome, wise, attractive
黄金脆皮鹅 (Golden Crispy Goose)	animal	Lady Gaga	celebrity	Lady Gaga dressed up like a goose in her "Applause" MV and "G.U.Y." MV.
轮胎 (tire)	product	胡锦涛 (Hu Jintao)	politics	锦湖 (Jin Hu) is Chinese translation of South Korean tire brand *Kumho*; which can be read in reverse as the first two characters of *Hu Jintao*'s name.
Angel	historical book	Air Force One	terrorism	Angel was used extensively for Air Force One throughout William Manchester's 1967 book, "The Death Of A President".
古月 (Gu Yue)	drama	胡锦涛 (Hu Jintao)	politics	the first character of 胡锦涛 can be decomposed into two characters 古月; and 古月 is a famous actor playing Mao Zedong. Both Mao and Hu acted as the former chairman of China.

Table 4: Domain Mapping Examples for Other Purposes

3 System Encoding and Decoding

3.1 Cipher for Encoding

Traditional text encryption techniques focus on alphabetic substitution or transposition based on lexical level (Franceschini and Mukherjee, 1996; Venkateswaran and Sundaram, 2010), synonyms (Chang and Clark, 2014) or image-adaptive public watermarking (Sun et al., 2008). Cipher systems can also be potentially developed and mutated to encode messages, including compare sophisticated ciphers such as historical ones (Knight et al., 2011) and simple mutations such as Leet[10] and Martian script[11]. Further strategies need to be developed to make them easy and fun for target human comprehension and widespread adoption, and difficult for automatic decoding.

3.2 Natural Language Generation for Encoding

Knight and Hatzivassiloglou (1995) and Langkilde and Knight (1998) lay the foundation for statistical natural language generation, and they recently apply these techniques to the generation of creative language:

(1) **Portmanteau neologism creation.** They fuse existing English words to create novel ones (Deri and Knight, 2015). The aim is to create an amusing new form that is understandable by a reader, e.g., *frenemy* for an entity that is both *friend* and *enemy*. Doing this well requires fusion at the phonetic level followed by an appropriate choice of spelling. Machines cannot yet process such created neologisms. This portmanteau generation approach (Deri and Knight, 2015) can be used to encode other types of neologisms

[10] https://en.wikipedia.org/wiki/Leet
[11] https://en.wikipedia.org/wiki/Martian_language

in both English and Chinese. We observe the words embedded are usually semantically and phonetically compatible, such as "*frenemy (friend + enemy)*". They are also terse, representative, expressive, interesting and easy to remember. For example, the following short phrases are created to refer to good men and bad men respectively: "暖男 *(warm + man)*" and "渣男 *(dirt + man)*".

(2) **Dynamic phrasebooks.** Tourists frequently carry phonetic phrasebooks that allow them to say things in a language they do not know. In (Shi et al., 2014), we developed a system that accepts text entered by a user (e.g., Chinese), translates the text (e.g., into English), then converts the translation into a phonetic spelling in the user's own orthography (e.g., Chinese). This system let users say anything they want, even if it is not in any phrasebook, using their own voice. For example, if a Chinese visitor to the United States wants to say 早上好 (meaning *Good morning*), they enter this phrase, and the system tells them to instead say 古德莫宁 (pronounced *gu-de-mo-ning*). This user is not required to know any English, but can still be understood by monolingual English speakers. This technique can be applied for encoding with the following two improvements: (1) replace human translation with machine translation (MT) to make the results more challenging to decode because it is difficult to recover from MT errors, and (2) polish the output by using common words or phrases in the user's own orthography, so the coded messages are more easily remembered and adopted. An encoded message is as follows.

- **Chinese**: 明天下午三点到鼓楼大街集合。
- **English translation**: Let's gather at the Bell Tower Street at 3pm tomorrow.
- **Pronounce English in Chinese phonetic system (pinyin)**: Laici galete ate de beier taoer sijute aite teli piaimu temoluo.
- **Code by spelling out pinyin**: 来此盖乐特爱特得贝尔套儿思聚特爱特特例皮埃姆特摩罗。

(3) **Poetry passwords.** (Greene et al., 2010) build the first statistical machine translation system to translate poetry. They subsequently apply poetry generation techniques to the problem of password security (Ghazvininejad and Knight, 2015). In this work, the machine first assigns a random 60-bit password to a user. Because the user cannot remember a random sequence of 0's and 1's, the machine converts the bit sequence into a more memorable iambic tetrameter couplet (e.g., *The legendary Japanese // subsidiaries overseas*). The mapping between bit sequences and poems is reversible, so the security of the randomly-assigned password is maintained.

Brennan et al. (2012) propose three methods to create adversarial passages: obfuscation, imitation, and translation. They find manual circumvention methods work very well, while automated translation methods are not effective. Potash et al. (2015) develop a *GhostWriter* system that can take a given artist's rap lyrics and generate similar yet unique lyrics. Other recent work detects word obfuscation in adversarial communication (Roussinov et al., 2007; Fong et al., 2008; Jabbari et al., 2008; Deshmukh et al., 2014; Agarwal and Sureka, 2015) using existing commonsense KBs such as ConceptNet (Agarwal and Sureka, 2015).

3.3 Entity Encoding and Decoding

Huang et al. (2013), Zhang et al. (2014) and Zhang et al. (2015) study a problem of encoding and decoding *entity morph*, which is a special case of coded name alias to hide the original entities for expressing strong sentiment or evading censorship in Chinese social media. They propose a variety of novel approaches to automatically encode proper and interesting morphs (Zhang et al., 2014), including Phonetic Substitution, Spelling Decomposition, Nickname Generation, Translation and Transliteration and Historical Figure Mapping, which can effectively pass decoding tests. They also develop an effective morph decoder, which can automatically identify and resolve entity morphs to their real targets (Huang et al., 2013; Zhang et al., 2015) and released manually annotated data sets for encoding[12] and decoding[13] experiments. They capture implicit attributes of entities using word embeddings. For example, they automatically encode 姚明 *(Yao Ming)* as 姚奇才 *(Yao Wizard)*. However, this simple approach based on general distributional semantics only receives 52% overall human satisfaction rate, which is significantly lower than that of human encoding methods (77%). They found that human methods are much better than automatic methods

[12] http://nlp.cs.rpi.edu/data/morphencoding.tar.gz
[13] http://nlp.cs.rpi.edu/data/morphdecoding.zip

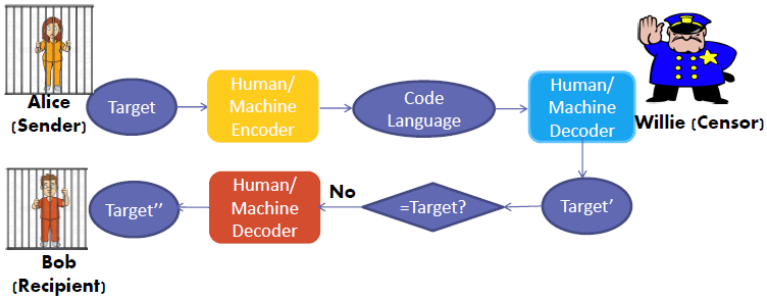

Figure 1: Encoding and Decoding as Prisoners Problem

at quickly picking up the best attributes for encoding, as well as connecting dots (seemingly irrelevant attributes) for decoding.

4 Remaining Problems

Many interesting challenges remain to be explored for effective creative language encoding.

Prisoners Problem: The first basic goal is to make sure the target human receiver (e.g., Bob in Figure 1) can successfully receive and decode the code words from the sender (Alice) by passing the censor (Willie). Creative language is used to communicate confidential information without encryption. Shannon's maxim "the enemy knows the system" does not always hold. There is usually no chance for the sender and the potential recipient to agree on a common code-book or secret key in advance. A smart encoder should go far beyond surface changes on spellings or pronunciations. It must be on-topic, containing clues for a human receiver to understand without specialized background knowledge.

Representative: Code words should be unique, expressive, discriminative, vivid to indicate what the target is well-known for, and easy to remember. For example, *Kimchi Country* is a good code name for South Korea. Terror code words often use the names of fruits and vegetables for weapons of similar shapes. (Lin et al., 2015) ranked the attributes in a typical human created knowledge base based on how well each attribute can distinguish entities. The top ranked ones include a person's social security number or an organization's website. However, such attributes will be easily caught by an automatic decoder. The good attributes do not necessarily appear very frequently either. They must be up-to-date for a certain burst event, topic, life snapshot, domain, or reputation during a certain time period.

Adoptable: Zhang et al. (2015) shows human created code words are considered as funnier (76% satisfaction rate) than machine created ones (46% satisfaction rate), which make them more impressive and adoptable for future use. For example, the character 牙 *(tooth)* in the code word for "*Luis Suarez*" in Table 2 is funny because it indicates his unusual habit of biting. The ultimate goal of encoding is to disseminate code words widely so they can become part of the new Internet language. Zhang et al. (2014) showed that popular code words tend to include unusual combinations of characters or words, negative sentiment and sarcasm, and they are usually simple, cute and easy to remember. However, a deeper and more comprehensive categorization and analysis on human created code words is required to formally extract their representative criteria for automated encoders to follow.

Time-sensitivity: Figure 2 shows the code frequency change for targets Bo Xilai at Sina Weibo. We can see the drops in the frequency of code words around sensitive dates (June 4, the anniversary of Tiananmen Square protests of 1989[14], etc). The censorship system is highly flexible. Certain code words may be blocked during politically-sensitive periods of the year, while others may be blacklisted for just a short time due to some contemporary relevant issue. Therefore, code words need to rapidly co-evolve with machine or human decoder over time, as some code words are discovered and blocked by censorship and newly created code words emerge. To fully automate human encoding approaches, a never-ending

[14]https://en.wikipedia.org/wiki/Tiananmen_Square_protests_of_1989

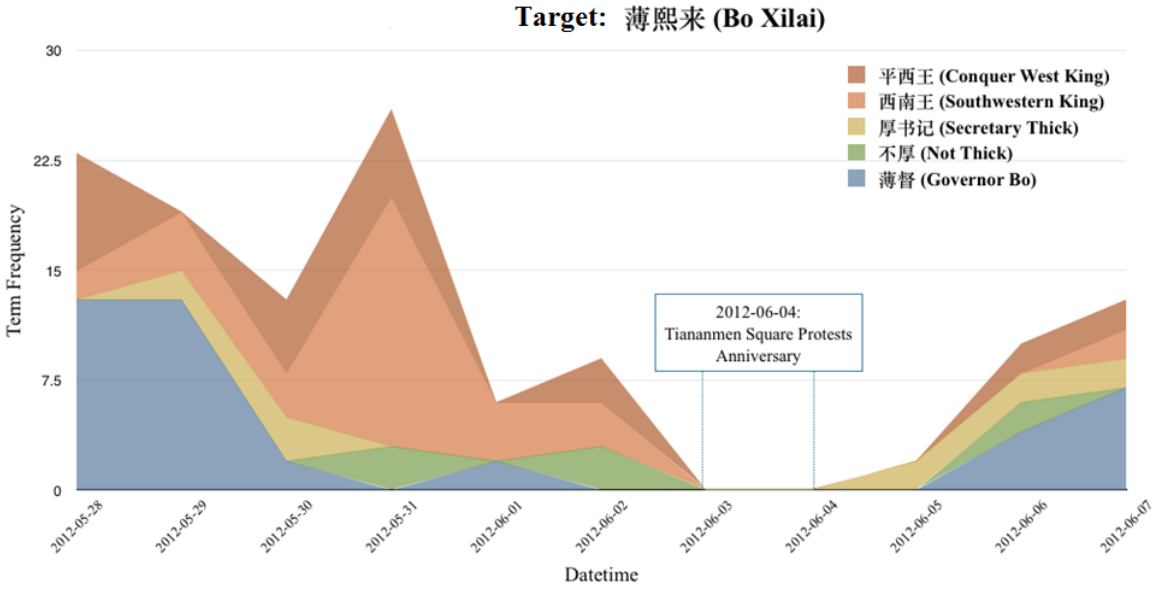

Figure 2: Code Frequency Change at Sina Weibo

learning component is needed to discover and select novel, salient and anomaly entities and events, update their profiles over time, rank and select the most discriminative characteristics of the targets for encoding.

Personalization: Most of the measures above are user-subjective. An ideal encoder should act as a friend instead of a servant or a co-worker. It should be a user-oriented, real-time, interactive encoding framework customized for both the sender and the receiver. The encoder should also incorporate the target reader's cultural background and psychological traits (gender, origin, education, moral value, power, belief, religion, social norm, current interests, culture, sense of humor, work intensity, etc), as recognized by user profile, social network information, target log history, previous posts, location, and time, into the encoder's short-term memory.

Multi-lingual: As we see from examples in English and Chinese, different encoding strategies can be developed for different writing systems. Another interesting research problems is to extend encoding methods to a wide range of languages which represent different writing systems and cover a large proportion of the world's population, including Latin, Cyrillic, Arabic, Devanagari, and Hangul scripts, as well as other languages in countries and regions with Internet censorship such as Arabic, Burmese, Farsi, Russian, Turkish, Uyghur, Urdu, and Vietnamese.

5 Conclusions

We summarized the current status and remaining challenges for a new and important research area of creative language encoding under censorship. Such encoding tools, if successful, will also be able to fast co-evolve over time with the semi-automatic censorship system' own evolutionary processes and ultimately defeat it. In the meanwhile it will bring new challenges and opportunities for adapting downstream natural language processing techniques to understand coded languages.

References

Swati Agarwal and Ashish Sureka. 2015. Using common-sense knowledge-base for detecting word obfuscation in adversarial communication. In *Proc. COMSNETS*.

David Bamman, Brendan O'Connor, and Noah A. Smith. 2012. Censorship and deletion practices in Chinese social media. *First Monday*, 17(3).

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security*, 15.

Ching-Yun Chang and Stephen Clark. 2014. Practical linguistic steganography using contextual synonym substitution and a novel vertex coding method. *Computational Linguistics*.

Le Chen, Chi Zhang, and Christo Wilson. 2013. Tweeting under pressure: Analyzing trending topics and evolving word choice on Sina Weibo. In *Proc. the first ACM conference on Online social networks*.

Aliya Deri and Kevin Knight. 2015. How to make a frenemy: Multitape FSTs for portmanteau generation. In *Proc. NAACL-HLT*.

Sonal N. Deshmukh, Ratnadeep R. Deshmukh, and Sachin N. Deshmukh. 2014. Performance analysis of different sentence oddity measures applied on Google and Google News repository for detection of substitution. *IRJES*, 3(3).

SzeWang Fong, Dmitri Roussinov, and David B. Skillicorn. 2008. Detecting word substitutions in text. *IEEE Transactions on Knowledge and Data Engineering*, 20(8).

Robert Franceschini and Amar Mukherjee. 1996. Data compression using encrypted text. In *Proc. IEEE Data Compression Conference*.

Marjan Ghazvininejad and Kevin Knight. 2015. How to memorize a random 60-bit string. In *Proc. NAACL-HLT*.

Erica Greene, Tugba Bodrumlu, and Kevin Knight. 2010. Automatic analysis of rhythmic poetry with applications to generation and translation. In *Proc. EMNLP*.

Ilana Heintz, Ryan Gabbard, Mahesh Srivastava, David Barner, Donald Black, Marjorie Friedman, and Ralph Weischedel. 2013. Automatic extraction of linguistic metaphors with LDA topic modeling. In *Proc. ACL Workshop on Metaphor in NLP*.

Chaya Hiruncharoenvate, Zhiyuan Lin, and Eric Gilbert. 2015. Algorithmically bypassing censorship on sina weibo with nondeterministic homophone substitutions. In *Proc. ICWSM*.

Hongzhao Huang, Zhen Wen, Dian Yu, Heng Ji, Yizhou Sun, Jiawei Han, and He Li. 2013. Resolving entity morphs in censored data. In *Proc. ACL*.

Sanaz Jabbari, Ben Allison, and Louise Guthrie. 2008. Using a probabilistic model of context to detect word obfuscation. In *Proc. LREC*.

Gary King, Jennifer Pan, and Margaret E. Roberts. 2014. Reverse-engineering censorship in China: Randomized experimentation and participant observation. *Science*, 6199(345).

Kevin Knight and Vasileios Hatzivassiloglou. 1995. Two-level, many-paths generation. In *Proc. ACL*.

Kevin Knight, Beata Megyesi, and Christiane Schaefer. 2011. The Copiale Cipher. In *Proc. ACL Workshop on Building and Using Comparable Corpora (BUCC)*.

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proc. COLING/ACL*.

Wen-Pin Lin, Matthew Snover, and Heng Ji. 2015. Unsupervised language-independent name translation mining from wikipedia infoboxes. In *Proc. EMNLP Workshop on Unsupervised Learning for NLP*.

Rada Mihalcea. 2007. Using wikipedia for automatic word sense disambiguation. In *Proc. HLT-NAACL*.

Maarten P. Mulder and Anton Nijholt. 2002. Humour research: State of the art.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Computing Surveys*, 41(10).

Xiaoman Pan, Taylor Cassidy, Ulf Hermjakob, Heng Ji, and Kevin Knight. 2015. Unsupervised entity linking with abstract meaning representation. In *Proc. the 2015 Conference of the North American Chapter of the Association for Computational Linguistics —Human Language Technologies (NAACL-HLT 2015)*.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2015. Ghostwriter: Using an LSTM for automatic rap lyric generation. In *Proc. EMNLP*.

Dmitri Roussinov, Sze Wang Fong, and David Skillicorn. 2007. Detecting word substitutions: Pmi vs. hmm. In *Proc. SIGIR*.

Xing Shi, Kevin Knight, and Heng Ji. 2014. How to speak a language without knowing it. In *Proc. ACL.*

Qiudong Sun, Wenxin Ma, Wenying Yan, and Hong Dai. 2008. Text encryption technique based on robust image watermarking. *Journal of Image and Graphics*.

Pedro Szekely, Craig A Knoblock, Jason Slepicka, Andrew Philpot, Amandeep Singh, Chengye Yin, Dipsy Kapoor, Prem Natarajan, Daniel Marcu, Kevin Knight, et al. 2015. Building and using a knowledge graph to combat human trafficking. In *The Semantic Web-ISWC 2015*. Springer.

Yulia Tsvetkov. 2013. Cross-lingual metaphor detection using common semantic features. In *Proc. ACL Workshop on Metaphor in NLP*.

R. Venkateswaran and Dr V. Sundaram. 2010. Information security: Text encryption and decryption with poly substitution method and combining the features of cryptography. *International Journal of Computer Applications*.

Zhimin Wang, Houfeng Wang, Huiming Duan, Shuang Han, and Shiwen Yu. 2006. Chinese noun phrase metaphor recognition with maximum entropy approach. In *Proc. CICLING*.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proc. ACL.*

Boliang Zhang, Hongzhao Huang, Xiaoman Pan, Sujian Li, Chin-Yew Lin, Heng Ji, Kevin Knight, Zhen Wen, Yizhou Sun, Jiawei Han, and Bulent Yener. 2013. Context-aware entity morph decoding. In *Proc. ACL.*

Boliang Zhang, Hongzhao Huang, Xiaoman Pan, Heng Ji, Kevin Knight, Zhen Wen, Yizhou Sun, Jiawei Han, and Bulent Yener. 2014. Be appropriate and funny: Automatic entity morph encoding. In *Proc. ACL.*

Boliang Zhang, Hongzhao Huang, Xiaoman Pan, Sujian Li, Chin-Yew Lin, Heng Ji, Kevin Knight, Zhen Wen, Yizhou Sun, Jiawei Han, and Bulent Yener. 2015. Context-aware entity morph decoding. In *Proc. ACL.*

Tao Zhu, David Phipps, Adam Pridgen, Jedidiah R. Crandall, and Dan S. Wallach. 2013. The velocity of censorship: High-fidelity detection of microblog post deletions. In *Proc. the 22nd USENIX Security Symposium*.