Detect Camouflaged Spam Content via StoneSkipping: Graph and Text Joint Embedding for Chinese Character Variation Representation

Zhuoren Jiang[1]*, Zhe Gao[2]*, Guoxiu He[3], Yangyang Kang[2], Changlong Sun[2], Qiong Zhang[2], Luo Si[2], Xiaozhong Liu[4]†

[1] School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China
[2]Alibaba Group, Hangzhou & Sunnyvale & Seattle, China & USA
[3]School of Information Management, Wuhan University, Wuhan, China
[4]School of Informatics, Computing and Engineering, Indiana University Bloomington, Bloomington, USA
jiangzhr3@mail.sysu.edu.cn,
{gaozhe.gz,yangyang.kangyy,qz.zhang,luo.si}@alibaba-inc.com,
guoxiu.he@whu.edu.cn, changlong.scl@taobao.com, liu237@indiana.edu

Abstract

The task of Chinese text spam detection is very challenging due to both glyph and phonetic variations of Chinese characters. This paper proposes a novel framework to jointly model Chinese variational, semantic, and contextualized representations for Chinese text spam detection task. In particular, a Variation Family-enhanced Graph Embedding (VFGE) algorithm is designed based on a Chinese character variation graph. The VFGE can learn both the graph embeddings of the Chinese characters (local) and the latent variation families (global). Furthermore, an enhanced bidirectional language model, with a combination gate function and an aggregation learning function, is proposed to integrate the graph and text information while capturing the sequential information. Extensive experiments have been conducted on both SMS and review datasets, to show the proposed method outperforms a series of state-of-the-art models for Chinese spam detection.

1 Introduction

Chinese orchestrates over tens of thousands of characters by utilizing their morphological information, e.g., pictograms, simple/compound ideograms, and phono-semantic compounds (Norman, 1988). Different characters, however, may share the similar glyph and phonetic "root". For instance, from glyph perspective, character "裸 (naked)" looks like "课 (course)" (homographs), while from phonetic viewpoint, it shares the similar pronunciation with "锣 (gong)" (homophones). The form of variations can also be compounded, for instance, "账 (account)"and "帐 (curtain)" have the similar structure and pronunciation (homonyms). Unfortunately, in the context

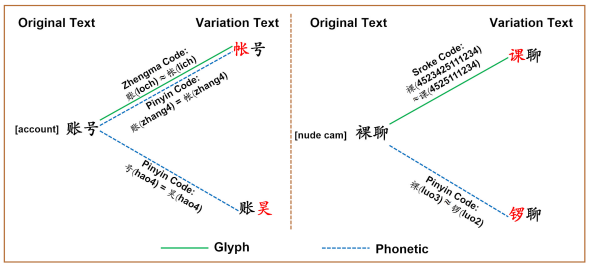

Figure 1: Character Variations in Chinese Spam Texts (the pinyin codes provide phonetic information and the stroke and Zhengma codes provide glyph information).

of spam detection, as shown in Figure 1, spammers are able to take advantage of these variations to escape from the detection algorithms (Jindal and Liu, 2007). For instance, in the e-commerce ecosystem, variation-based Chinese spam mutations thrive to spread illegal, misleading, and harmful information[1]. In this study, we propose a novel problem - Chinese Spam Variation Detection (CSVD), a.k.a. investigating an effective Chinese character embedding model to assist the classification models to detect the variations of Chinese spam text, which needs to address the following key challenges.

Diversity: the variation patterns of Chinese characters can be complex and subtle, which are difficult to generalize and detect. For instance, in the experimental dataset, one Chinese character can have 297 (glyph and phonetic) variants averagely and 2,332 maximally. The existing keyword based spam detection approaches, e.g., (Ntoulas et al., 2006), can hardly address this problem. **Sparseness, Zero-shot, and Dynamics**: when competing with classification models, spammers are constantly creating new Chinese characters combinations for spam texts (that can be a

*These two authors contributed equally to this research.
†Corresponding author

[1]More detailed information can be found in the experiment section.

"zero/few shot learning" problem (Socher et al., 2013)). The labelling cost can be inevitably high in such dynamic circumstance. Data driven approaches, e.g., (Zhang et al., 2015), will perform poorly for unseen data. **Camouflage**: with the common cognition knowledge of Chinese and the contextual information, users are able to consume the spam information, even when some characters in the content are intentionally mutated into their similar variations (Spinks et al., 2000; Shu and Anderson, 1997). However, the variation-based spam text are highly camouflaged for machines. It is important to propose a novel Chinese character representation learning model that can synthesize character variation knowledge, semantics, and contextualized information.

To address these challenges, we propose a novel solution, **S**tone**S**kipping (**SS**) model to enable Chinese variation representation learning via graph and text joint embedding. SS is able to learn the Chinese character variation knowledge and predict the new variations not appearing in the training set by utilizing sophisticated heterogeneous graph mining method. For a piece of text (a character sequence), with the proposed model, each candidate character can probe character variation graph (like stone bouncing cross the water surface), and explore its glyph and phonetic variation information (like the ripples caused by the stone hitting the water). Algorithmically, a Variation Family-enhanced Graph Embedding (VFGE) algorithm is proposed to extract the heterogeneous Chinese variation knowledge while learning the (local) graph representation of a Chinese character along with the (global) representation of the latent variation families. Finally, an enhanced bidirectional language model, with a combination gate function and an aggregation learning function, is proposed to comprehensively learn the variation, semantic, and sequential information of Chinese characters. To the best of our knowledge, this is the first work to use graph embedding to learn the heterogeneous variation knowledge of Chinese characters for spam detection.

The major contributions of this paper can be summarized as follows:

1. We propose an innovative CSVD problem, in the context of text spam detection, to address the diversity, sparseness, and text camouflage problems.

2. A novel joint embedding SS model is proposed to learn the variational, semantic, and contextual representations of Chinese characters. SS is able to predict unseen variations.

3. A Chinese character variation graph is constructed for encapsulating the glyph and phonetic relationships among Chinese characters. Since the graph can be potentially useful for other NLP tasks, we share the graph/embeddings to motivate further investigation.

4. Through the extensive experiments on both SMS and review datasets[2], we demonstrate the efficacy of the proposed method for Chinese spam detection. The proposed method outperforms the state-of-the-art models.

2 Related Work

Neural Word Embeddings. Unlike traditional word representations, low-dimensional distributed word representations (Mikolov et al., 2013; Pennington et al., 2014) are able to capture in-depth semantics of text content. More recently, ELMo (Peters et al., 2018) employed learning functions of the internal states of a deep bidirectional language model to generate the character embeddings. BERT (Devlin et al., 2018) utilized bidirectional encoder representations from transformers (Vaswani et al., 2017) and achieved improvements for multiple NLP tasks. However, all the prior models only focused on learning the context, whereas the text variation was ignored. Moreover, CSVD problem can be different from other NLP tasks: the intentional character mutations and unseen variations (zero-shot learning (Socher et al., 2013)) can threaten prior models' performances.

Chinese Word and Sub-word Embeddings. A number of studies explored Chinese representation learning methodologies. CWE (Chen et al., 2015) learned the character and word embeddings to improve the representation performance. GWE (Su and Lee, 2017) introduced the features extracted from the images of traditional Chinese characters. JWE (Yu et al., 2017) used deep learning to generate character embedding based on an extended radical collection. Cw2vec (Cao et al., 2018) investigated Chinese character as a sequence of n-gram stroke order to generate its embedding. Although these models had considered the nature of

[2]In order to help other scholars reproduce the experiment outcome, we will release the datasets via GitHub (https://github.com/Giruvegan/stoneskipping)

Chinese characters, they only utilized glyph features while the phonetic information was ignored. In CSVD problem, the forms of variations can be heterogeneous, and a single kind of features cannot cover all mutation patterns. More importantly, all these models are not designed for spam detection, and the task-oriented model should be able to highlight the most important variations for spam text.

Graph Embedding. Graph (a.k.a. information network) is a natural data structure for characterizing the multiple relationships between the objects. Recently, multiple graph embedding algorithms are proposed to learn the low dimensional feature representations of vertexes in graphs. DeepWalk (Perozzi et al., 2014) and Node2vec (Grover and Leskovec, 2016) are random walk based models. LINE (Tang et al., 2015) modeled 1st and 2nd order graph neighbourhood. Meanwhile, metapath2vec++ (Dong et al., 2017) was designed for heterogeneous graph embedding with human defined metapath rules. HEER (Shi et al., 2018) is a recent state-of-the-art heterogeneous graph embedding model. Though the techniques utilized in these models are different, most existing graph embedding models focus more on local graph structure representation, e.g., modelling of a fixed-size graph neighbourhood. CSVD problem requires graph embedding conducted from a more global perspective, to characterize comprehensive variation patterns.

Spelling Correction. Spelling correction may serve as an alternative to address CSVD problem, e.g., using dictionary-based (Yeh et al., 2014) or language model-based method (Yu and Li, 2014) to restore the content variations to their regular format. However, because spammers intentionally mutate the spam text to escape from the detection model, training data sparseness and dynamics may challenge this approach.

3 StoneSkipping Model

Figure 2 depicts the proposed SS model. There are three core modules in SS: a **Chinese character variation graph** to host the heterogeneous variation information; a **variation family-enhanced graph embedding** for Chinese character variation knowledge extraction and graph representation learning; an **enhanced bidirectional language model** for joint representation learning. In the remaining of this section, we will introduce them in detail.

3.1 Chinese Character Variation Graph

A Chinese character variation graph[3] can be denoted as $G = (C, R)$. C denotes the Chinese character set, and each character is represented as a vertex in G. R denotes the variation relation (edge) set, and edge weight is the similarity of two characters given the target relation (variation) type. To accurately characterize both phonetic and glyph information of Chinese character, we utilize three different encoding methods:

Pinyin system provides phonetic-based information, which is widely used for representing the pronunciations of Chinese characters (Chen and Lee, 2000). In this system, each Chinese character has one syllable which consists of three components: an initial (consonant), a final (vowel), and a tone. There are four types of tones in Modern Standard Mandarin Chinese. Different tones with the same syllable can have different meanings. For instance, the pinyin code of "裸 (naked)" is "luo3" and "锣 (gong)" is 'luo2". The pinyin-based variation similarity is calculated based on their pinyin syllables with tones[4].

Stroke is a basic glyph pattern for writing Chinese character (Cao et al., 2018). All Chinese characters are written in a certain stroke order and can be represented as a stroke code, e.g., the stroke code of "裸 (naked)" is "4523425111234" and '课 (course)" is "4525111234". The stroke-based variational similarity is calculated based on *longest common substring* and *longest common subsequence* metrics[4].

Zhengma is another important means for glyph character encoding, which encodes character at radical level (Yu et al., 2017). For instance, the Zhengma code of "裸 (naked)" is "WTKF" and '课 (course)" is "SKF". The Zhengma-based variational similarity is calculated based on the *Jaccard Index* metric[4].

Unlike previous works (Cao et al., 2018; Yu et al., 2017) only employ one kind of glyph-based information, we utilize two different glyph patterns (stroke and Zhengma) to encode the Chinese character. Because these two patterns can characterize Chinese characters from different internal structural levels, and complement each other

[3]https://github.com/Giruvegan/stoneskipping

[4]Because of the space limitation, the detailed operations of relation generation will be provided on https://github.com/Giruvegan/stoneskipping.

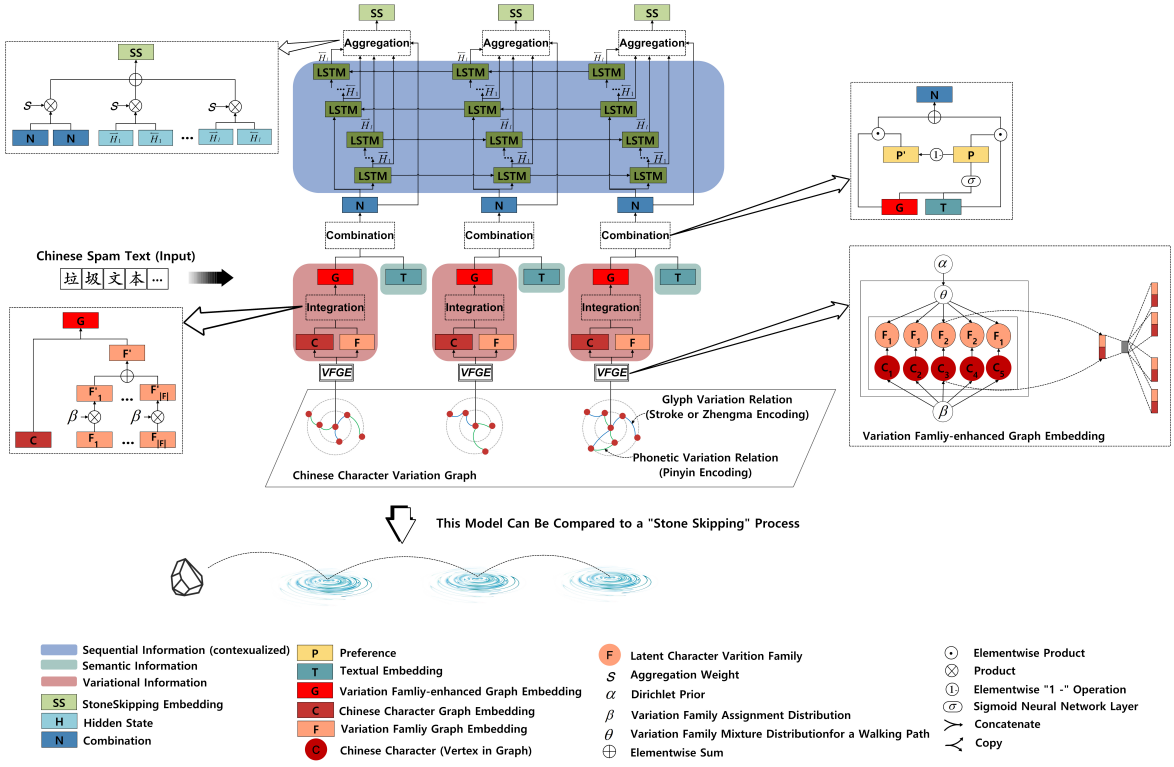

Figure 2: An Illustration of "StoneSkipping" Framework

to enable an enhanced glyph representation learning. Furthermore, the pinyin encoder provides phonetic information. The constructed character variation graph integrates these three kinds of variation relations, which can be significant for camouflaged spam detection.

3.2 Variation Family-enhanced Graph Embedding

While the variation graph can provide comprehensive knowledge of Chinese character variations, efforts need to be made to address these two problems: (1) the variation patterns can be very flexible, and the compounded (long-range) variation information transfer may exist. Therefore, short-range (local) graph information, e.g., character vertex's neighbors, may be insufficient for spam detection. Meanwhile, it is impractical to exhaust all the possible variation patterns. (2) To oblige users to consume the text content, spammers cannot make the variation patterns to be too complex/confusing, they usually focus on the most sensitive words in a spam message. Hence, some random infrequent variation patterns could be "noisy" for CSVD while polluting the detection outcomes.

Latent Character Variation Family. In this study, we propose a VFGE model to address these problems. As depicted in Figure 2, in VFGE model, we introduce a set of latent variables "*character variation family*" $F = \{F_1, ..., F_{|F|}\}$ at a graph schema (global) level to capture the critical information for spam detection. Each F_i is defined as *a distribution of characters*, which aims to estimate the global frequent variation dependencies in G. By learning F, VFGE is able to highlight the useful variations, eliminate the noisy patterns, and predict the unseen variation forms w.r.t. the spam detection task.

Random Walk based Character Family Representation Co-Learning . VFGE is a random walk based graph embedding model, and we employ a hierarchical random walk strategy (Jiang et al., 2018b) on G to generate the optimized walking paths (character vertex sequences) for each character. The model can sample the most possible variation context vertexes for each character. Based on generated walking paths, VFGE executes the following two processes iteratively:

(1) **Family Assignment**. By leveraging both local context and global family distributions, we assign a discrete family for each character vertex in a particular walking path to form a character-family pair $\langle C, F \rangle$. As shown in Figure 2, we assume dif-

ferent walking paths tend to emerge various character variation patterns which can be represented as mixtures over latent variation families. Given a character C_i in a $path$, C_i has a higher chance to be assigned to a dominant family F_i. The assignment probability can be calculated as:

$$\begin{aligned} & Pr(F_i|C_i, path) \\ \propto & Pr(C_i, F_i, path) \\ = & Pr(path)Pr(F_i|path)Pr(C_i|F_i) \end{aligned} \tag{1}$$

As depicted in Figure 2, α is the parameter of the Dirichlet prior on the per-path family distributions ($Pr(path)$); β is the family assignment distribution ($Pr(C|F)$); and θ is the family mixture distribution for a walking path ($Pr(F|path)$). The distribution learning can be considered as a Bayesian inference problem, and we use Gibbs sampling (Porteous et al., 2008) to address this problem.

(2) **Character-Family Representation Co-Learning**. Given the assigned character-family pairs, the proposed method aims to obtain the representations of character C and latent variation family F by mapping them into a low-dimensional space $\mathbb{R}^d$ (d is a parameter specifying the number of dimensions). Motivated by (Liu et al., 2015), we propose a novel representation learning method to optimize characters and families ***separately*** and ***simultaneously***.

The objective is defined to maximize the following log probability:

$$\mathcal{L} = \max_f \sum_{C_i \in C} \sum_{C_j \in \mathbb{N}(C_i)} log Pr(\langle C_j, F_j \rangle | \mathbf{C_i^{F_i}}) \tag{2}$$

We use $f(\cdot)$ as the embedding function. $\mathbf{C_i} = f(C_i)$ represents the character graph embedding and $\mathbf{F_i} = f(F_i)$ represents the family graph embedding. $\mathbf{C_i^{F_i}}$ denotes the concatenation of $\mathbf{C_i}$ and $\mathbf{F_i}$, whereas $\mathbb{N}(C_i)$ is C_i's neighborhood (context). As Figure 2 shows, the feature representation learning method is an upgraded version of the skip-gram architecture. Compared with merely using the target vertex C_i to predict context vertexes in original skip-gram model (Mikolov et al., 2013), the proposed approach employs the character-family pair $\langle C_i, F_i \rangle$ to predict context character-family pairs. From variation viewpoint, character vertex's context will encapsulate both local (vertex) and global (variation family) information. Hence, the learned representations are able to comprehensively preserve the variation information in G.

$Pr(\langle C_j, F_j \rangle | \mathbf{C_i^{F_i}})$ is modeled as a softmax function:

$$Pr(\langle C_j, F_j \rangle | \mathbf{C_i^{F_i}}) = \frac{exp(\mathbf{C_j^{F_j}} \cdot \mathbf{C_i^{F_i}})}{\sum_{C_k \in C} exp(\mathbf{C_k^{F_k}} \cdot \mathbf{C_i^{F_i}})} \tag{3}$$

Stochastic gradient ascent is used for optimizing the model parameters of f. Negative sampling (Mikolov et al., 2013) is applied for optimization efficiency. Note that, the parameters of each character embedding and family embedding are shared over all the character-family pairs, which, as suggested in (Liu et al., 2015), can address the training data sparseness problem and improve the representation quality.

Family-enhanced Embedding Integration. As shown in Figure 2, the family-enhanced character graph embedding can be calculated as:

$$\mathbf{G_i} = \left[\mathbf{C_i}, \sum_{F_j \in F} Pr(F_j|C_i)\mathbf{F_j} \right] \tag{4}$$

where $\mathbf{G_i}$ is family-enhanced graph embedding for C_i, and $[\cdot]$ is concatenating operation. $Pr(F_j|C_i)$ can be inferred from family assignment distribution β.

3.3 Enhanced Bidirectional Language Model

As shown in Figure 2, SS model utilizes an enhanced bidirectional language model to jointly learn variation, semantic and contextualized representation of Chinese character.

Combination Gate Function. This gate function is utilized for combining the variation and semantic representations, which is the input function for bidirectional language model. The formulations of the gate function are listed as follows:

$$\begin{aligned} \mathbf{P} &= \sigma(\mathbf{W}_P \cdot [\mathbf{G}, \mathbf{T}] + \mathbf{b}_P) \\ \mathbf{N} &= (\mathbf{P} \odot \mathbf{T}) + ((\mathbf{1} - \mathbf{P}) \odot \mathbf{G}) \end{aligned} \tag{5}$$

$\mathbf{P} \in \mathbb{R}^d$ is the preference weights for controlling the contributions from $\mathbf{G} \in \mathbb{R}^d$ (variation graph embedding) and $\mathbf{T} \in \mathbb{R}^d$ (Skip-Gram textual embedding). $\mathbf{W}_P \in \mathbb{R}^{2d \times d}$. $\mathbf{N} \in \mathbb{R}^d$ is the combination representation. $\odot$ is elementwise product, and $+$ is elementwise sum.

Aggregation Learning Function. With the combination representation $\mathbf{N}$ as input, we train a bidirectional language model for capturing the sequential information. There could be multiple

Group	Model	SMS		Review	
		Accuracy	F1 Score	Accuracy	F1 Score
Text	**Skipgram** (Mikolov et al., 2013)	0.807	0.765	0.693	0.560
	GloVe (Pennington et al., 2014)	0.732	0.637	0.707	0.600
	ELMo (Peters et al., 2018)	0.786	0.747	0.755	0.647
Chinese	**CWE** (Chen et al., 2015)	0.751	0.674	0.780	0.726
	GWE (Su and Lee, 2017)	0.505	0.426	0.778	0.718
	JWE (Yu et al., 2017)	0.770	0.707	0.738	0.646
	Cw2vec (Cao et al., 2018)	0.800	0.753	0.724	0.618
Graph	**DeepWalk** (Perozzi et al., 2014)	0.836	0.804	0.738	0.638
	LINE (Tang et al., 2015)	0.821	0.783	0.764	0.695
	Node2vec (Grover and Leskovec, 2016)	0.835	0.802	0.792	0.736
	M2V$_{\mathbf{Max}}$ (Dong et al., 2017)	0.838	0.807	0.790	0.740
	HEER (Shi et al., 2018)	0.723	0.617	0.771	0.708
Correction	**Pycorrector** (Yu and Li, 2014)	0.782	0.727	0.688	0.549
Comparison	**SS**$_{\mathbf{Graph}}$	0.839	0.827	0.812	0.756
	SS$_{\mathbf{Naive}}$	0.849	0.825	0.811	0.757
	SS$_{\mathbf{Original}}$	**0.851**	**0.832**	**0.854**	**0.822**

Table 1: Chinese Text Spam Detection Performance Comparison of Different Models

layers of forward and backward LSTMs in bidirectional language model. For k_{th} character, $\overrightarrow{\mathbf{H}}_l^k$ is the forward LSTM unit's output for layer l, where $l = 1, 2, ..., L$, and $\overleftarrow{\mathbf{H}}_l^k$ is the output of the backward LSTM unit.

The output **SS** embedding is learned from an aggregation function, which aims to aggregate the intermediate layer representations of the bidirectional language model and the input embedding $\mathbf{N}$. For k_{th} character, if we denote $\mathbf{H}_0^k = [\mathbf{N}^k, \mathbf{N}^k]$ (self concatenation), and $\mathbf{H}_l^k = [\overleftarrow{\mathbf{H}}_l^k, \overrightarrow{\mathbf{H}}_l^k]$, the output can be:

$$\mathbf{SS^k} = \omega\Big(\underbrace{s_0 \mathbf{H}_0^k}_{\text{(Variational \& Semantic)}} + \underbrace{\sum_{l=1}^{L} s_l \mathbf{H}_l^k}_{\text{Contextualized}} \Big) \quad (6)$$

where ω is the scale parameter, and s_l is a weight parameter for the combination of each layer, which can be learned through the training process. Similar aggregation operation has been proven useful to model the contextualized word representation (Peters et al., 2018).

4 Experiment

4.1 Dataset and Experiment Setting

Dataset[5]. In Table 2, we summarize the statistics of the two real-world spam datasets (in Chinese). One is a SMS dataset, the other is a product review dataset. Both datasets were manually labeled (spam or regular labels) by professionals. False advertising and scam information are the most common forms of spam information for SMS dataset, while abuse information dominates review spam dataset.

Dataset	Part	All	Spam	Normal
SMS	Train	48,884	23,891	24,993
	Test	48,896	23,891	25,005
Review	Train	37,299	17,299	20,000
	Test	37,299	17,299	20,000

Table 2: Statistics of Two Chinese Spam Text Datasets

In the constructed variation graph, there are totally 25,949 Chinese characters (vertexes) and 7,705,051 variation relations. For all the variation relations, there are 1,508,768 pinyin relations (phonetic), 373,803 stroke relations (glyph), and 5,822,480 Zhengma relations (glyph).

Experimental Set-up. We validated the proposed model in Chinese text spam detection task. In order to simulate the "diversity", "sparseness" and " zero-shot" problems under real business scenarios, we made a challenging restriction on the training and testing sets, i.e., the character variations were only included in testing set, and all samples in training set were using the original characters.

For the proposed SS model, we utilized the following setting: layers of LSTMs: 2; dimension of hidden (output) state in LSTM: 128; dimension of pre-trained character text embedding: 128; dimen-

[5]https://github.com/Giruvegan/stoneskipping

Character	Text Skipgram		Chinese Cw2vec		Graph VFGE		Proposed model SS	
运(move)	C	捷(prompt)	C	捷(prompt)	G P	云(cloud)	S C	转(transmit)
	C	站(stop)	C	站(stop)	G P	纭(numerous)	G P	芸(weed)
	C	客(guest)	S C	输(transport)	G	坛(altar)	G P	云(cloud)
惊(shock)	S C	讶(surprised)	S C	讶(surprised)	G P	景(view)	S C	慌(flurried)
	S C	愕(startled)	S C	撼(shake)	G	晾(dry)	G	琼(jade)
	S C	吓(scare)	S C	愕(startled)	G	谅(forgive)	G S C	悚(afraid)

G: Glyph; P: Phonetic; S: Semantic; C :Context

Table 3: Case Study: given the target character, we list the top 3 similar characters from each algorithm. The characters are selected from a frequently used candidate character set whose size is 8238.

sion of VFGE embedding: 128; batch size: 64; Dropout: 0.1. For training VFGE embedding[6], the walk length was 80, the number of walks per vertex was 10. These parameters were adopted in (Peters et al., 2018; Jiang et al., 2018a; Perozzi et al., 2014; Grover and Leskovec, 2016). The variation family number[7] was 500. SS model was pretrained for parameter initialization as suggested in (Peters et al., 2018).

Baselines and Comparison Groups. We chose 13 strong baseline algorithms, from text or graph viewpoints, to comprehensively evaluate the performance of the proposed method.

General Textual Based Baselines: **Skip-gram** (Mikolov et al., 2013), **GloVe** (Pennington et al., 2014), and **ELMo** (Peters et al., 2018).

Chinese Specific Textual Based Baselines: **CWE** (Chen et al., 2015), **GWE** (Su and Lee, 2017), **JWE** (Yu et al., 2017), and **Cw2vec** (Cao et al., 2018).

Graph Embedding Based Baselines: **DeepWalk** (Perozzi et al., 2014), **LINE** (Tang et al., 2015), **Node2vec** (Grover and Leskovec, 2016), **Metapath2vec++** (Dong et al., 2017), and **HEER** (Shi et al., 2018). We applied this group of baselines on constructed Chinese character variation graph to get graph based character embeddings. Specifically, **Metapath2vec++** required a human-defined metapath scheme to guide the random walks. We tried 4 different metapaths for this experiment:(1) $\mathbf{M2V_P}$ (only walking on pinyin (phonetic) relations); (2) $\mathbf{M2V_S}$ (only walking on stroke (glyph) relations); (3) $\mathbf{M2V_Z}$ (only walking on Zhengma (glyph) relations); (4) $\mathbf{M2V_C}$ (alternately walking on glyph and phonetic relations). We reported the best results from these four metapaths, denoted as $\mathbf{M2V_{Max}}$.

Spelling Correction Baseline: **Pycorrector**[8] based on n-gram language model (Yu and Li, 2014).

Comparison Group: we compared the performances of several variants of the proposed method in order to highlight our technical contributions. There were 3 comparison groups conducted. $\mathbf{SS_{Graph}}$: we only used VFGE graph embedding. $\mathbf{SS_{Naive}}$: we simply concatenated VFGE graph embedding and skip-gram textual embedding (a naive version). $\mathbf{SS_{Original}}$: the proposed SS model.

For a fair comparison, the dimension[9] of all embedding models was 128. A single layer of CNN classification model[10] was used for spam detection task.

4.2 Experiment Result and Analysis

The text spam detection task performances of different models were reported in Table 1. Based on the experiment results, we had the following observations:

SS vs. Baselines. (1) $\mathrm{SS_{Original}}$ outperformed the baseline models for all evaluation metrics on both datasets, which indicated the proposed SS model can effectively address the CSVD problem. (2) On review dataset, the leading gap between $\mathrm{SS_{Original}}$ and other baselines was greater. A possible explanation was that, the review spam text

[6]For the experiment fairness, all the random walk based graph embedding baselines shared the same parameters with VFGE.

[7]Based on the parameter sensitive analysis, the proposed method was not very sensitive to number of variation families.

[8]https://github.com/shibing624/pycorrector

[9]The initial dimension of $\mathrm{SS_{Naive}}$ and $\mathrm{SS_{Original}}$ is 256, so we used a fully connected layer to reduce its dimension to 128.

[10]The filter sizes of CNN is 3, 4, 5, and the filter number is 128, dropout ratio is 0.1.

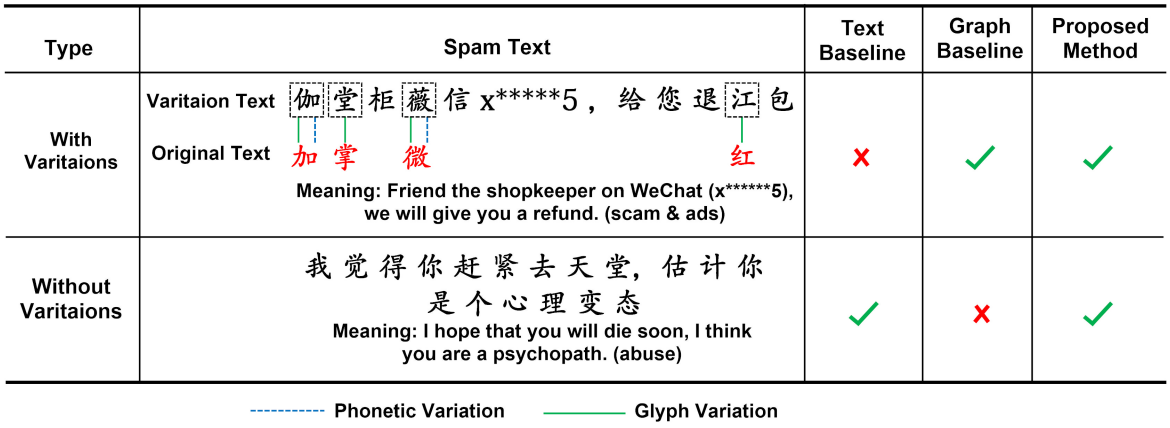

Figure 3: Two typical examples for CSVD task

usually had richer content and more complex variation patterns than SMS spam text. Therefore, a good variation representation model may have certain advantages.

Chinese vs. General. (1) Compared to classical textual embedding models (Skipgram and GloVe), the Chinese embedding models showed their advantages, especially on review dataset. This result indicated that the characteristic knowledge of Chinese can help to detect spam text. (2) ELMo was able to learn both the semantic and contextualized information, and it achieved a good performance in text baseline group.

Graph vs. Text. Generally, the graph based baselines outperformed the textual based baselines (including general and Chinese). This observation indicated: (1) the variation knowledge of Chinese character can be critical for CSVD problem. (2) The proposed character variation graph can provide critical information for Chinese character representation learning. (3) Compared to other graph based baselines, $\mathrm{SS}_{\mathbf{Graph}}$ was superior, which proved the effectiveness of VFGE algorithm, and the proposed variation family can characterize and predict useful variation patterns for CSVD problem.

Chinese Character Encodings. (1) In Chinese textual embedding baseline group, JWE (radical based) and Cw2vec (stroke based) didn't perform well, which indicated employing a single kind of glyph-based information can be insufficient for Chinese variation representation learning. Similarly, in graph based baseline group, the performances of $\mathrm{M2V}_{\mathbf{P}}$, $\mathrm{M2V}_{\mathbf{S}}$ and $\mathrm{M2V}_{\mathbf{Z}}$ (employed only one encoding relation on the constructed graph) were still unsatisfactory. The results revealed that an individual encoding method cannot comprehensively encode a character, we should consider various kinds of variation information simultaneously. (2) The performance of $\mathrm{M2V}_{\mathbf{C}}$ (integrated all relations based on a predefined metapath pattern) was still inferior. This result indicated a human-defined rule cannot effectively integrate all relationships in a complex graph.

Representation vs. Spelling Correction. Pycorrector performed poorly in experiment, and other baselines outperformed this approach, which proved the spelling correction method is not capable for CSVD problem.

Variants of SS model. For variants of the proposed method, the results showed that (1) by combining the semantic and sequential information, the task performances can improve; (2) simply concatenating graph and text embeddings cannot generate a satisfactory joint representation. (3) The proposed SS model can successfully capture the variation, semantic, and sequential information for character representation learning.

4.3 Case Study

. To gain an insightful understanding regarding the variation representation of the proposed method, we conduct qualitative analysis by performing the case studies of character similarities. As shown in Table 3, for exemplary characters, the most similar characters, based on skipgram embedding (general textual based baseline), are all semantically similar or/and context-related. Meanwhile, based on Cw2vec embedding (most recent Chinese embedding baseline), all similar characters for target characters are also semantically similar or/and context-related. Unsurprisingly, for each target character, all similar characters based on

VFGE model (best performed graph embedding model), are glyph and phonetic similar characters. The proposed SS model can achieve a comprehensive coverage from variation, semantic and context viewpoints. For instance, in its top 3 similar characters for "运(move)", "转(transmit)" is a semantic and context similar character, and "云(cloud)" is a glyph and phonetic similar character. Furthermore, SS model can capture complicated compound similarity between Chinese characters, for instance, "悚(afraid)" is a glyph, semantic, and context similar character for "惊(shock)". This also explains why SS model performs well to address the CSVD problem.

Figure 3 depicts two typical examples in the experimental datasets. For the spam text with variations, spammers used character variations to create camouflaged expressions. For instance, using glyph variation "江(river)" to replace "红(red)", and glyph-phonetic compound variation "薇(osmund)" to replace "微(micro)". The classical text embedding models may fail to identify this kind of spam texts. With the mining of character variation graph, the graph based approaches can be successful to capture these changes. For spam text without variations, classification models need more semantic and contextual information, and the text based methods can be suitable for this kind of spam texts. The proposed SS model is able to detect both two kinds of spam texts effectively, and experiment results proved SS can successfully model Chinese variational, semantic and contextualized representations for CSVD task.

5 Conclusion

In this paper, we propose a StoneSkipping model for Chinese spam detection. The performance of the proposed method is comprehensively evaluated in two real world datasets with challenging experimental setting. The results of experiments show that the proposed model significantly outperforms a number of state-of-the-art methods. Meanwhile, the case study empirically proves that the proposed model can successfully capture the Chinese variation, semantic, and contextualized information, which can be essential for CSVD problem. In the future, we will investigate more sophisticated methods to improve SS's performance, e.g., enable self-attention mechanism for contextualized information modelling.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61876003, 81971691), the China Department of Science and Technology Key Grant (2018YFC1704206), and Fundamental Research Funds for the Central Universities (18lgpy62).

References

Shaosheng Cao, Wei Lu, Jun Zhou, and Xiaolong Li. 2018. cw2vec: Learning chinese word embeddings with stroke n-gram information. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5053–5061.

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huan-Bo Luan. 2015. Joint learning of character and word embeddings. In *IJCAI*, pages 1236–1242.

Zheng Chen and Kai-Fu Lee. 2000. A new statistical approach to chinese pinyin input. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 135–144. ACM.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864. ACM.

Zhuoren Jiang, Liangcai Gao, Ke Yuan, Zheng Gao, Zhi Tang, and Xiaozhong Liu. 2018a. Mathematics content understanding for cyberlearning via formula evolution map. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 37–46. ACM.

Zhuoren Jiang, Yue Yin, Liangcai Gao, Yao Lu, and Xiaozhong Liu. 2018b. Cross-language citation recommendation via hierarchical representation learning on heterogeneous graph. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 635–644. ACM.

Nitin Jindal and Bing Liu. 2007. Review spam detection. In *Proceedings of the 16th international conference on World Wide Web*, pages 1189–1190. ACM.

Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2015. Topical word embeddings. In *AAAI*, pages 2418–2424.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Jerry Norman. 1988. *Chinese*. Cambridge University Press.

Alexandros Ntoulas, Marc Najork, Mark Manasse, and Dennis Fetterly. 2006. Detecting spam web pages through content analysis. In *Proceedings of the 15th international conference on World Wide Web*, pages 83–92. ACM.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 701–710. ACM.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 2227–2237.

Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2008. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577. ACM.

Yu Shi, Qi Zhu, Fang Guo, Chao Zhang, and Jiawei Han. 2018. Easing embedding learning by comprehensive transcription of heterogeneous information networks. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2190–2199. ACM.

Hua Shu and Richard C Anderson. 1997. Role of radical awareness in the character and word acquisition of chinese children. *Reading Research Quarterly*, 32(1):78–89.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943.

John A Spinks, Ying Liu, Charles A Perfetti, and Li Hai Tan. 2000. Reading chinese characters for meaning: The role of phonological information. *Cognition*, 76(1):B1–B11.

Tzu-ray Su and Hung-yi Lee. 2017. Learning chinese word representations from glyphs of characters. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 264–273.

Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*, pages 1067–1077. International World Wide Web Conferences Steering Committee.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Jui-Feng Yeh, Yun-Yun Lu, Chen-Hsien Lee, Yu-Hsiang Yu, and Yong-Ting Chen. 2014. Chinese word spelling correction based on rule induction. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 139–145.

Jinxing Yu, Xun Jian, Hao Xin, and Yangqiu Song. 2017. Joint embeddings of chinese words, characters, and fine-grained subcharacter components. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 286–291.

Junjie Yu and Zhenghua Li. 2014. Chinese spelling error detection and correction based on language model, pronunciation, and shape. In *Proceedings of The Third CIPS-SIGHAN Joint Conference on Chinese Language Processing*, pages 220–223.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.