Pronunciation-Enhanced Chinese Word Embedding

Qinjuan Yang¹ · Haoran Xie² · Gary Cheng³ · Fu Lee Wang⁴ · Yanghui Rao¹

Received: 11 November 2020 / Accepted: 4 February 2021 / Published online: 22 February 2021
© The Author(s) 2021

Abstract

Chinese word embeddings have recently garnered considerable attention. Chinese characters and their sub-character components, which contain rich semantic information, are incorporated to learn Chinese word embeddings. Chinese characters can represent a combination of meaning, structure, and pronunciation. However, existing embedding learning methods focus on the structure and meaning of Chinese characters. In this study, we aim to develop an embedding learning method that can make complete use of the information represented by Chinese characters, including phonology, morphology, and semantics. Specifically, we propose a pronunciation-enhanced Chinese word embedding learning method, where the pronunciations of context characters and target characters are simultaneously encoded into the embeddings. Evaluation of word similarity, word analogy reasoning, text classification, and sentiment analysis validate the effectiveness of our proposed method.

Keywords Chinese embedding · Pronunciation · Chinese characters · Sentiment analysis

Introduction

Word embedding (also known as distributed word representation) denotes a word as a real-valued and low-dimensional vector. In recent years, it has attracted significant attention and has been applied to many natural language processing (NLP) tasks, such as sentiment classification [1–5],

sentence/concept-level sentiment analysis [6–10], question answering [11, 12], text analysis [13, 14], named entity recognition [15, 16], and text segmentation [17, 18]. For instance, in some prior studies, word embeddings in a sentence were summed and averaged to obtain the probability of each sentiment [19]. An increasing number of researchers have used word embeddings as the input of neural networks in sentiment analysis tasks [20]. This approach can facilitate neural networks to encode semantic and syntactic information of words into embedding and subsequently place the words with the same or similar meaning close to each other in a vector space. Among the existing methods, the continuous bag-of-words (CBOW) method and continuous skip-gram (skip-gram) method are popular because of their simplicity and efficiency to learn word embeddings from large corpora [21, 22].

Due to its success in modeling English documents, word embedding has been applied to Chinese text. In contrast to English where a word is the basic semantic unit, characters are considered as the smallest meaningful units in Chinese and are called morphemes in morphology [23]. A Chinese character may form a word by itself or, in most occasions, be a part of a word. Chen et al. [24] integrated context words with characters to improve the learning of Chinese word embedding. From the perspective of morphology, a character can be further decomposed into sub-character components, which contain rich semantic

✉ Gary Cheng
chengks@eduhk.hk

Qinjuan Yang
yangqinj@mail2.sysu.edu.cn

Haoran Xie
hrxie2@gmail.com

Fu Lee Wang
pwang@ouhk.edu.hk

Yanghui Rao
raoyangh@mail.sysu.edu.cn

¹ School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China

² Department of Computing and Decision Sciences, Lingnan University, Tuen Mun, New Territories, Hong Kong, China

³ Department of Mathematics and Information Technology, The Education University of Hong Kong, Tai Po, Hong Kong, China

⁴ School of Science and Technology, The Open University of Hong Kong, Ho Man Tin, Kowloon, Hong Kong, China

or phonological information. With the help of the internal structural information of Chinese characters, many studies tried to improve the learning of Chinese word embeddings by using radicals [25, 26], sub-word components [27], glyph features [28], and strokes [29].

These methods enhanced the quality of Chinese word embeddings in terms of two distinct perspectives: morphology and semantics. In particular, they explore the semantics of characters within different words through the internal structure of characters. However, we argue that such information is insufficient to capture semantics because a Chinese character may represent different meanings in different words and the semantic information cannot be completely drawn from their internal structures. As shown in Fig. 1, “亠” is the radical of “道”, but it can merely represent the first meaning. “道” can be decomposed into “亠” and “首”, which are still semantically related to the first meaning. The stroke n-gram feature “首” appears to have no relevance to any semantics. Although Chen et al. [24] proposed to address this issue by incorporating a character’s positional information in words and learning position-related character embeddings, their method cannot identify distinct meanings of a character. For instance, “道” can be used at the beginning of multiple words but with distinct meanings, as illustrated in Fig. 1.

Chinese characters represent a combination of pronunciation, structure, and meaning, which correspond to phonology, morphology, and semantics in linguistics. However, the aforementioned methods consider the structure and meaning only. Phonology describes “the way sounds function within a given language to encode meaning” [30]. In Chinese language, the pronunciation of Chinese words and characters is marked by Pinyin, which stems from the official Romanization system for Standard Chinese [31]. The Pinyin system includes five tones, namely, flat tone, rising tone, low tone, falling tone, and neutral

character	meaning	word
dào 道	road	道路(road, way) 铁道(railway)
	rule	道理(principle) 道德(morality)
	Taoism	道教(Taoism)

radical	components	stroke n-gram

Fig. 1 Illustrative example of radical, components, and stroke n-gram of character “道”

pinyin	mā	má	mǎ	mà	ma
character	妈	麻	马	骂	吗
meaning	mother	hemp	horse	scold	final interrogative particle

Fig. 2 Example of five characters with five tones from syllable *ma*

tone. Figure 2 shows that the syllable, “ma”, can produce five different characters with different tones.

In contrast to English characters where one character has only one pronunciation, many Chinese characters can have two or more pronunciations. These are called Chinese polyphonic characters. Each pronunciation can usually refer to several different meanings. In modern Standard Chinese, one-fifth of the 2400 most common characters have multiple pronunciations¹.

Let us take the Chinese polyphonic character, “长”, as an example. As Fig. 3 shows, the character has two pronunciations: “cháng” and “zhǎng”, and each has multiple semantics for distinct words. It is possible to uncover different meanings of the same character based on its pronunciation.

Although a character may represent several meanings with the same pronunciation, we can identify its meaning by using the phonological information of other characters within the same word. This is possible because the semantic information of the character is determined when collocated with other pronunciations. Moreover, the position problem in character-enhanced word embedding (CWE) and multi-granularity embedding (MGE) can be addressed because the position of a character in a word can be determined when its pronunciation is considered. For example, when “长” is combined with “(generation)”, they can only form the word, “长辈”. In this case, “长” represents “old”. In another case, “长” and “年 (year)” can produce two different words, “年长” and “长年”. They can be distinguished by the pronunciation of “长”, where the former is “zhǎng” and the latter is “cháng”. Hence, it would be helpful to incorporate the pronunciation

character	pinyin	semantic	example
长	cháng	long	长久(for a long time), 长期(long-term), 长远(long-term)
		distance	长度(length)
	zhǎng	increase	增长(rise), 助长(encourage)
		grow	生长(growth), 成长(grow up), 长大(grow)
		old	年长(older), 长辈(elder)
		chief	厨师长(chef), 市长(mayor)

Fig. 3 Chinese polyphonic character “长”

¹ http://pinyin.info/chinese_characters/

of characters to learn Chinese word embeddings and improve the ability to capture polysemous words.

In this study, we propose a method named pronunciation-enhanced Chinese word embedding (PCWE), which makes full use of the information represented by Chinese characters, including phonology, morphology, and semantics. Pinyin, which is the phonological transcription of Chinese characters, is combined with Chinese words, characters, and sub-character components as context inputs of PCWE. Word similarity, word analogy reasoning, text classification, and sentiment analysis tasks are evaluated to validate the effectiveness of PCWE.

Proposed Method

In this section, we describe the details of PCWE, which makes full use of phonological, internal structural, and semantic features of Chinese characters based on CBOW [22]. We do not use the skip-gram method because there are insignificant differences between CBOW and skip-gram but CBOW runs slightly faster [32]. PCWE uses context words, context characters, context sub-characters, and context pronunciation to predict the target word.

We denote the training corpus as D , vocabulary of words as W , character set as C , sub-character set as S , phonological transcription set as P , and size of the context window as T . As can be seen in Fig. 4, PCWE attempts to maximize the sum of four log-likelihoods of conditional probabilities for the target word, w_i , given the average of individual context vectors:

$$L(w_i) = \sum_{k=1}^4 p(w_i | h_{ik}), \quad (1)$$

where h_{i1}, h_{i2}, h_{i3} , and h_{i4} are the compositions of context words, context characters, context sub-characters, and context phonological transcriptions, respectively. Let $v_{w_i}, v_{c_{w_i}}, v_{s_{w_i}}$, and $v_{p_{w_i}}$ denote the vector of word w_i , character c_i , sub-character s_i , and phonological transcription p_i , respectively. $\hat{v}_{w_i}$ is the predictive vector of the target word, w_i . The conditional probability is defined as

$$p(w_i | h_{ik}) = \frac{\exp(h_{ik}^T \hat{v}_{w_i})}{\sum_{n=1}^N \exp(h_{ik}^T \hat{v}_{w_n})}, k = 1, 2, 3, 4, \quad (2)$$

where h_{i1} is the average of the vectors of the context words:

$$h_{i1} = \sum_{-T \leq j \leq T, j \neq 0} v_{w_{i+j}}. \quad (3)$$

Similarly, h_{i2}, h_{i3} , and h_{i4} are the averages of the vectors of the characters, sub-characters, and pronunciations in the context, respectively.

A similar objective function was used in joint learning word embedding (JWE) [27]. However, there are

Fig. 4 Structure of PCWE, where w_i is the target word, w_{i-1} and w_{i+1} are the context words, c_{i-1} and c_{i+1} represent the context characters, s_{i-1} and s_{i+1} indicate the context sub-characters, p_{i-1} and p_{i+1} denote the context pronunciation, and p_i is the pronunciation of w_i

differences between JWE and PCWE. PCWE integrates the phonological information of words and characters with words and sub-character components to jointly learn Chinese word embedding, whereas JWE incorporates only words, characters, and sub-character component features. In other words, PCWE makes full use of context information from the perspective of both phonology and morphology, whereas JWE merely uses the morphological features. PCWE utilizes the pronunciation of target characters, but JWE does not.

Experiments

We tested our method in terms of word similarity, word analogy reasoning, text classification, and sentiment analysis. Qualitative case studies were also conducted.

Datasets and Experiment Settings

We employed the Chinese Wikipedia Dump², as our training corpus. In preprocessing, we used THULAC³ for segmentation and parts-of-speech-tagging. Pure digits and non-Chinese characters were removed. Finally, we obtained a 1GB training corpus with 169,328,817 word tokens and 613,023 unique words. We used the lists of radicals and sub-character components in [27], comprising 218 radicals, 13,253 components, and 20,879 characters.

We crawled 411 Chinese Pinyin spellings without tones from the online Chinese Dictionary website⁴. Since the Pinyin system includes five tones, a total of 2055 Pinyin spellings were collected. We adopted HanLP⁵ to transfer Chinese words into Pinyin. Finally, we obtained 3,289,771 Chinese words and Pinyin pairs.

PCWE was compared with CBOW [22]⁶, CWE [24]⁷, position-based CWE (CWE+P), MGE [26]⁸, and JWE [27]⁹. The same parameter settings were applied to all methods for comparison. We set the context window size to 5, embedding dimension to 200, and training iteration to 100. We used 10-word negative sampling and set the initial learning rate to 0.025 and the subsampling parameter to 10^{-4} for process of the optimization. Words with a frequency of less than 5 were ignored during training.

Word Similarity

This task evaluates the ability to capture the semantic relativity between two embeddings. We adopted two Chinese word-similarity datasets, Wordsim-240 and Wordsim-297, provided by Chen et al. [24]¹⁰ for evaluation. Both datasets contained Chinese word pairs with human-labeled similarity scores. There were 240 and 297 pairs of Chinese words in Wordsim-240 and Wordsim-297, respectively. However, there were 8 words in Wordsim-240 and 10 words in Wordsim-297 that did not appear in the training corpus. We removed these words to generate Wordsim-232 and Wordsim-287.

The cosine similarity of two word embeddings was computed to measure the similarity score of word pairs.

Table 1 Results on word similarity

Method	Wordsim-232	Wordsim-287
CBOW	0.5322	0.5746
CWE	0.5138	0.6022
CWE+P	0.5075	0.5960
MGE	0.4635	0.5231
JWE	0.5706	0.6541
PCWE	0.5544	0.6743

We calculated the Spearman correlation [33] between the similarity scores computed by using the word embeddings and the human-labeled similarity scores. Higher values of the Spearman correlation denote better word embedding for capturing semantic similarity between words. The evaluation results are listed in Table 1.

The results show that PCWE outperforms CBOW, CWE, CWE+P, and MGE on the two word similarity datasets. This indicates that the combination of morphological, semantic, and phonological features can exploit deeper Chinese word semantic and phonological information than that exploited by other methods. In addition, we find that PCWE can yield competitive results when compared with CWE+P. This verifies the benefits of exploiting phonological features over position features to reduce the ambiguity of Chinese characters within different words. Although PCWE performs better than all the baselines on Wordsim-287, JWE achieves the best results on Wordsim-232. A possible reason for this could be that Wordsim-287 contains more Chinese polyphonic characters such as “行” (háng, xíng)) and “中”.

Word Analogy Reasoning

This task estimates the effectiveness of word embeddings to reveal linguistic regularities between word pairs. Given three words, a , b , and c , the above task aims to explore a fourth word, d , such that a to b is similar to c to d . We used 3CosAdd [34] and 3CosMul [35] to determine the nearest word, d . We employed the analogy dataset provided by Chen et al. [24] that contained 1127 Chinese word tuples. They were categorized into three types: capitals of countries (677 tuples), state/provinces of cities (175 tuples), and family words (240 tuples). The training corpus covered all the words in the analogous dataset. We used accuracy as the evaluation metric and the results are listed in Tables 2 and 3.

From the results of word analogy reasoning with the 3CosAdd measure function, it can be found that PCWE achieves the second best performance, whereas JWE performs the best. Nevertheless, the embedding representations learned by PCWE have better analogies for Capital and City types when computing with the 3CosMul

² <https://dumps.wikimedia.org/zhwiki/>

³ https://thulac.thunlp.org/sendMessage_v1_1

⁴ <http://xh.5156edu.com/pinyin.html>

⁵ <https://github.com/hankcs/HanLP/>

⁶ <https://code.google.com/p/word2vec/>

⁷ <https://github.com/Leonard-Xu/CWE>

⁸ We implement MGE base on the code of CWE.

⁹ <https://github.com/HKUST-KnowComp/JWE>

¹⁰ <https://github.com/Leonard-Xu/CWE/tree/master/data>

Table 2 Results of word analogy reasoning measured in 3CosAdd

Method	Total	Capital	City	Family
CBOW	0.6699	0.7622	0.7200	0.4081
CWE	0.7687	0.8744	0.88	0.4338
CWE+P	0.7865	0.8641	0.8857	0.5294
MGE	0.6388	0.7696	0.8343	0.1875
JWE	0.8301	0.8953	0.9200	0.6103
PCWE	0.8185	0.8922	0.9029	0.5809

function. This could be attributed to the fact that the words in these categories rarely consist of Chinese polyphonic characters, and meanwhile the morphological features provide sufficient semantic information to identify similar word pairs. To verify this hypothesis, we used a statistical technique to analyze these datasets. Only 3.46% of words in all tuples contained Chinese polyphonic characters that can affect the semantics. By contrast, the ratios for Wordsim-232 and Wordsim-287 were 15.73% and 10.98%, respectively. It can also be found that the results of PCWE are better than those of CWE and CWE+P, suggesting the success of leveraging the compositional internal structure and phonological information.

Text Classification

Text classification is widely used to evaluate the effectiveness of word embeddings in NLP tasks [36]. We adopted the *Fudan* dataset, which contained documents on 20 topics, for training¹¹ and testing¹². Following [29], we selected 12,545 (6424 for training and 6121 for testing) documents on five topics: environment, agriculture, economy, politics, and sports. We averaged the embeddings of the words that were present in the documents as the features of the documents.

Table 3 Results of word analogy reasoning measured in 3CosMul

Method	Total	Capital	City	Family
CBOW	0.7536	0.7563	0.6971	0.3971
CWE	0.7687	0.8538	0.8629	0.4338
CWE+P	0.7731	0.8479	0.9029	0.5294
MGE	0.7415	0.8171	0.1654	0.6139
JWE	0.8185	0.8847	0.9086	0.5956
PCWE	0.8176	0.8996	0.9143	0.5515

We also trained a classifier with LIBLINEAR¹³ [37]. The accuracy of different methods is given in Table 4.

As shown in the table, all the methods achieve an accuracy of over 94% and our method performs the best. This is because the distinct semantics of characters with different pronunciations are captured by our method. For example, for documents on the topic of economy, the word “银行 (bank)” with the polyphonic character “行” is used frequently and its pronunciation can contribute more to the accuracy than its subcomponents. Therefore, the PCWE outperforms other baselines.

Sentiment Analysis

Sentiment analysis [38–41] is one of the most important applications of word embedding [19, 42, 43]. It can also be used to evaluate the effectiveness of our method. With the popularity of Chinese texts, an increasing number of researchers have conducted Chinese sentiment analysis using Chinese word embedding methods [8, 44]. We chose the Hotel Reviews dataset¹⁴ in this experiment and analyzed the sentiment of each document to evaluate our word-embedding method. This dataset contained 2000 positive reviews and 2000 negative reviews. We randomly divided the dataset into two parts. One was the training set, which accounted for 90% of the data. The other one was the testing set, which contained the remaining 10% of the data. We built a bidirectional long short-term memory to perform the sentiment analysis task [45, 46]. The average accuracy of different methods is shown in Table 5.

It can be observed that PCWE performs significantly better than the employed word embedding baselines when applied to sentiment analysis. In the process of analyzing characters with more than one pronunciation, PCWE can capture different meanings of different pronunciations. For example, the negative reviews contain several words such as “假四星酒店 (fake four-star hotel)”. In this word, the

Table 4 Accuracy of text classification

Method	Accuracy (%)
CBOW	94.5
CWE	94.2
CWE+P	94.5
MGE	94.4
JWE	94.7
PCWE	94.9

¹¹ http://download.csdn.net/download/github_36326955/9747927

¹² http://download.csdn.net/download/github_36326955/9747929

¹³ <https://github.com/cjlin1/liblinear>

¹⁴ https://www.aitechclub.com/data-detail?data_id=29

Table 5 Accuracy of sentiment analysis

Method	Accuracy (%)
CBOW	79.2
CWE	80.1
CWE+P	81.9
MGE	80.4
JWE	83.6
PCWE	86.5

polyphonic character “假” is the core of the meaning. The pronunciation of the polyphonic character is considerably more important than its sub-character component for Chinese sentimental analysis.

Case Study

In addition to validating the benefit of using the phonetic information in Chinese characters to improve the word embedding quality, we performed qualitative analysis by conducting case studies to present the most similar words to certain target words.

Figure 5 illustrates the top-10 similar words to two target words identified by each method. The first example of the target word is “强壮 (qiáng zhuàng, strong)”. This target word includes a Chinese polyphonic character, “强”, which is used to describe “strong man” or “strong power”. The majority of the top-ranked words generated by CBOW have

no relevance to the target word, such as “鳍状肢 (flipper)” and “尾巴 (tail)”. This results from the fact that CBOW only incorporated context information. CWE exploits many words related to the characters constituting the target word “(强” or “壮)”. This verifies the idea of CWE to jointly learn embeddings of words and characters. However, CWE generates the word “瘦弱 (emaciated),” which represents the opposite meaning to the target word. MGE is the worst method which discovers words that are not semantically related to the target word, such as “主密码 (master password)” and “短尾蝠 (Mystacina tuberculata)”. By contrast, JWE yields words that are correlated with the target word, except “聪明 (smart).” As the best method, PCWE identifies words that are highly semantically relevant to the target word. In addition, only PCWE can generate the word “大块头 (big man),” which is related to the target word. Overall speaking, PCWE can effectively capture semantic relevance, as it combines the comprehensive information in characters from the perspectives of morphology, typography, and phonology.

The other example is “朝代 (dynasty),” which contains a Chinese polyphonic character, “朝.” It refers to an emperor’s reign or a certain emperor of a pedigree. CWE generated irrelevant words such as “分封制 (the system of enfeoffment)” and “典章制度 (ancient laws and regulations),” which are under the theme of laws and institutions. CWE also identifies irrelevant words under the topic of calendar, such as “大统历 (The Grand Unified Calendar)” and “统元历 (The Unified Yuan Calendar).” As

target word	CBOW	CWE	MGE	JWE	PCWE
强壮(strong)	健壮(fitness)	健壮(fitness)	主密码(master password)	健壮(fitness)	强健(robust)
	结实(burly)	粗壮(sturdy)	强健(robust)	强健(robust)	健壮(fitness)
	强健(robust)	强健(robust)	金眸(The Golden Eyes Ball)	粗壮(sturdy)	粗壮(sturdy)
	鳍状肢(flipper)	身强体壮(well-built)	短尾蝠(Mystacina tuberculata)	高大(tall)	结实(burly)
	尾巴(tail)	壮健(strong and healthy)	健硕(strong and muscular)	结实(burly)	健硕(strong and muscular)
	矮胖(tubby)	壮硕(imposing)	锐利(sharp)	体格(physique)	壮健(strong and healthy)
	前肢(fore limb)	坚韧(tough)	体格(physique)	壮硕(imposing)	壮硕(imposing)
	短尾巴(short tail)	肥壮(stout)	马羚亚科(Hippotraginae)	聪明(smart)	高大(tall)
	身躯(body)	瘦弱(emaciated)	波塞东龙(Sauroposeidon)	壮健(strong and healthy)	体格(physique)
	长尾巴(long tail)	凶猛(ferocious)	腕力(wrist power)	勇猛(bold and powerful)	大块头(big man)
朝代(dynasty)	历朝(successive dynasties)	历朝历代(successive reigns and dynasties)	藩属国(vassal state)	封建王朝(feudal dynasties)	历朝(successive dynasties)
	各朝(each dynasty)	历代(past dynasties)	历朝(successive dynasties)	王朝(dynasty)	封建王朝(feudal dynasties)
	隋唐(The Sui and Tang dynasties)	历朝(successive dynasties)	萍踪侠影录(Stories of the Wandering Hero)	历朝(in the past dynasties)	王朝(dynasty)
	分封制(the system of enfeoffment)	两朝(two dynasties)	类书(reference books with material taken from various sources and arranged according to subjects)	历朝历代(successive reigns and dynasties)	各朝(each dynasty)
	前朝(the proceeding dynasty)	诸侯国(vassal state)	年号(the title of an emperor's reign)	妃嫔(imperial concubine)	历朝历代(successive reigns and dynasties)
	历朝历代(successive reigns and dynasties)	大统历(The Grand Unified Calendar)	盛衰(prosperity and decline)	史书(historical records)	年号(the title of an emperor's reign)
	典章制度(ancient laws and regulations)	封建王朝(feudal dynasties)	封建王朝(feudal dynasties)	各朝(each dynasty)	君主(monarch)
	历代(past dynasties)	列朝(successive dynasties)	叶榆县(The Yeyu County)	两朝(two dynasties)	皇朝(dynasty)
	明清(Ming and Qing Dynasties)	各朝(each dynasty)	相权(prime minister's power)	历代(past dynasties)	历代(past dynasties)
	封建王朝(feudal dynasties)	统元历(The Unified Yuan Calendar)	嫡庶(son born of the legal wife and son born of a concubine)	君主(monarch)	汉朝(The Han Dynasty)

Fig. 5 Case study for semantically related words. Given the target word, the top 10 similar words identified by each method are listed

character	pinyin
长	zhǎng , cháng, fān
当	dāng , dàng, zhǎng
将	jiàng , jiāng, pōu
藏	cáng , zàng, gá
朝	cháo , nǎng, jiǒng
强	qiáng , kài, wá
林	lín , xú, tái
智	zhì , miǎo, zhuāi

Fig. 6 Case study for related Chinese characters and pronunciations. Given a Chinese character, the top 3 related pronunciations are listed. The correct Pinyin is in boldface

same as the first example, the words found by MGE appear to have no correlation with the target word. In contrast, the majority of the similar words generated by JWE are semantically correlated with the target word, but the words “妃嫔 (imperial concubine)” and “史书 (historical records)” present no relevance to the target word. For PCWE, all generated words are highly semantically relevant to the target word.

The singular inclusion of context word information in CBOW leads to the generation of contextual words instead of semantically related words. The learning processes of CWE, MGE, and JWE involve characters, radicals, and internal structures, resulting in the limitation of identifying similar words with the same characters, radicals, and internal structure but with no semantic relevance to the target word. Our proposed method PCWE can exploit more complete information of characters by considering their phonological features.

In addition to demonstrating the effectiveness of our method to encode phonological information, we conducted a case study to exploit the relationship between characters and pronunciations. We first analyzed the effectiveness of our method in capturing Chinese polyphonic characters and

listed the related Pinyin to the given character, as shown in Fig. 6. From the results, it is found that PCWE identifies correct pronunciations for every character. It is also found that our method does not find all possible pronunciations for the characters because some pronunciations are rarely used. Figure 7 depicts the relevant characteristics for the given Pinyin. The characters exploited for the first three Pinyin are all pronounced as expected. For the fourth Pinyin “shān,” only the character “山” can be correctly identified. Although the other characters are not pronounced as the target Pinyin, they are all semantically related to hills or mountains. The same situation occurs for the last Pinyin “.” The results show that our method can encode the polyphonic features of characters and explore semantically relevant characters.

Related Work

Most methods designed for word embedding learning are based on CBOW or skip-gram [21, 22] because of their effectiveness and efficiency. However, they regard words as basic units and ignore the rich internal information within words. Therefore, many methods have been exploited to improve word embeddings by incorporating morphological information. Bojanowski et al. [47] and Wieting et al. [48] proposed methods for learning word embeddings with character n-grams. Goldberg and Avraham [49] proposed a method to capture both semantic and morphological information, where each word is composed of vectors of linguistic properties. Cao and Lu [50] introduced a method based on a convolutional neural network that learns word embeddings with character 3-gram, root/affix, and inflections. Cotterell and Schütze [51] proposed a method that learns morphological word embeddings with morphological annotated data by extending the log-bilinear method. Bhatia et al. [52] introduced a morphological prior distribution to improve word embeddings. Satapathy et al. [53] developed a systematic approach called PhonSenticNet to integrate phonetic and string features of words for concept-level sentiment analysis.

In recent years, methods specifically designed for the Chinese language where characters are treated as the basic

pinyin	yí	shī	lín	shān	shuǐ
	夷(yí, raze)	师(shī, teacher)	林(lín, forest)	山(shān, hill)	水(shuǐ, water)
	移(yí, move)	诗(shī, poetry)	琳(lín, beautiful jade)	麓(lù, foot of a hill)	泗(qiú, swim)
character	遗(yí, lose)	失(shī, lose)	临(lín, look down from above)	岳(yuè, high mountain)	饮(yǐn, drink)
	宜(yí, fitting)	施(shī, apply)	霖(lín, long-continue drain)	圉(gù, small hill that is sheer all round but flat on the top)	淹(yān, drown)
	仪(yí, appearance)	狮(shī, lion)	邻(lín, neighbor)	岩(yán, cliff)	粼(lín, clear water)

Fig. 7 Case study for related Chinese characters and pronunciations. Given a Pinyin, the top 5 related Chinese characters are listed

semantic units have been studied. Chen et al. [24] utilized characters to augment Chinese word embeddings and proposed a CWE method to jointly learn Chinese characters and word embeddings. Yang and Sun [54] considered the semantic knowledge of characters when combining them with context words. There are also methods that incorporate internal morphological features of characters to enhance the learning of Chinese word embeddings. With a structure similar to that of [54], Xu et al. [55] combined characters with their semantic similarity with words. Yin et al. [26] proposed MGE by extending CWE with the radicals of target words. Considering that radicals cannot fully uncover the semantics of characters, Yu et al. [27] proposed the JWE method to utilize context words, context characters, and context sub-characters. Li et al. [25] combined the context characters and their respective radical components as inputs to learn character embeddings. Shi et al. [32] decomposed the contextual character sequence into a radical sequence to learn radical embedding. Su and Lee [28] enhanced Chinese word representation using character glyph features to learn from the bitmaps of characters. Cao et al. [29] learned Chinese word embeddings by exploiting stroke-level information. More recently, Peng et al. [56] have proposed two methods to encode phonetic information and integrate the representations for Chinese sentiment analysis. Unlike the said studies, our method incorporates multiple features of Chinese characters in terms of morphology, semantics, and phonology.

Conclusion and Future Work

In this study, we propose a method named PCWE to learn Chinese word embeddings. It incorporates various features of Chinese characters from the perspectives of morphology, semantics, and phonology. Experimental results of word similarity, word analogy reasoning, text classification, sentiment analysis, and case studies validate the effectiveness of our method. In the future, we plan to explore comprehensive strategies for modeling phonological information by integrating our PCWE with other resources [7] or methods [6], and exploit the core idea of the proposed method to address other state-of-the-art NLP tasks, including sentiment analysis [57], reader emotion classification [58], review interpretation [59], empathetic dialogue systems [60], end-to-end dialogue systems [61], and stock market prediction [62, 63].

Acknowledgements This research was supported by Research Grants Council of Hong Kong SAR, China (UGC/FDS16/E01/19), General Research Fund (No. 18601118) of Research Grants Council of Hong Kong SAR, China, One-off Special Fund from Central and Faculty Fund in Support of Research from 2019/20 to 2021/22 (MIT02/19-20) of The Education University of Hong Kong, Hong Kong, HKIBS Research Seed Fund 2019/20 (190-009), the Research Seed Fund

(102367), and LEO Dr David P. Chan Institute of Data Science of Lingnan University, Hong Kong. We are grateful to Xiaorui Qin for her work on sentiment analysis experiments.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

Informed consent Informed consent was not required as no human or animals were involved.

Human and animal rights This article does not contain any studies with human or animal subjects performed by any of the authors.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Xiong S, Lv H, Zhao W, Ji D. Towards twitter sentiment classification by multi-level sentiment-enriched word embeddings. *Neurocomputing*. 2018;275:2459–66.
2. Yu L, Wang J, Lai KR, Zhang X. Refining word embeddings using intensity scores for sentiment analysis. *IEEE/ACM Trans. Audio, Speech & Language Processing*. 2018 26(3), 671–681.
3. Majumder N, Poria S, Peng H, Chhaya N, Cambria E, Gelbukh A. Sentiment and sarcasm classification with multitask learning. *IEEE Intell Syst*. 2019;34(3):38–43.
4. Lo SL, Cambria E, Chiong R, Cornforth D. A multilingual semi-supervised approach in deriving singlish sentic patterns for polarity detection. *Knowl-Based Syst*. 2016;105:236–47.
5. Basiri ME, Nemati S, Abdar M, Cambria E, Acharya UR. ABCDM: an attention-based bidirectional CNN-RNN deep model for sentiment analysis. *Futur Gener Comput Syst*. 2021;115:279–94.
6. Cambria E, Fu J, Bisio F, Poria S. Affectivespace 2: Enabling affective intuition for concept-level sentiment analysis. In: *Proceedings of the 29th AAAI Conference on Artificial Intelligence*. 2015. pp. 508–514.
7. Peng H, Cambria E. Csentinet: A concept-level resource for sentiment analysis in Chinese language. In: *Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing*. 2017. pp. 90–104.
8. Peng H, Cambria E, Zou X. Radical-based hierarchical embeddings for Chinese sentiment analysis at sentence level. In: *Proceedings of the 30th International FLAIRS Conference*. 2017. pp. 347–352.
9. Peng H, Ma Y, Li Y, Cambria E. Learning multi-grained aspect target sequence for Chinese sentiment analysis. *Knowl-Based Syst*. 2018;148:167–76.

10. Poria S, Cambria E, Winterstein G, Huang GB. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowl-Based Syst.* 2014;69:45–63.
11. Shen Y, Rong W, Jiang N, Peng B, Tang J, Xiong Z. Word embedding based correlation model for question/answer matching. In: *Proceedings of the 31st AAAI Conference on Artificial Intelligence.* 2017. pp. 3511–3517.
12. Ghosal D, Majumder N, Gelbukh AF, Mihalcea R, Poria S. COSMIC: commonsense knowledge for emotion identification in conversations. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings.* 2020. pp. 2470–2481.
13. Fast, E., Chen, B., Bernstein, M.S.: Lexicons on demand: Neural word embeddings for large-scale text analysis. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence.* 2017. pp. 4836–4840.
14. Le TMV, Lauw HW. Semantic visualization for short texts with word embeddings. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence.* 2017. pp. 2074–2080.
15. Shijia E, Xiang Y. Chinese named entity recognition with character-word mixed embedding. In: *Proceedings of 2017 ACM on Conference on Information and Knowledge Management.* 2017. pp. 2055–2058.
16. Zhong X, Cambria E, Rajapakse JC. Named entity analysis and extraction with uncommon words. 2018. CoRR abs/1810.06818
17. Zhou H, Yu Z, Zhang Y, Huang S, Dai X, Chen J. Word-context character embeddings for Chinese word segmentation. In: *Proceedings of 2017 Conference on Empirical Methods in Natural Language Processing.* 2017. pp. 760–766.
18. Ma J, Hinrichs EW. Accurate linear-time Chinese word segmentation via embedding matching. In: *Proceedings of the 51rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing.* 2015. pp. 1733–1743.
19. Tang D, Wei F, Yang N, Zhou M, Liu T, Qin B. Learning sentiment-specific word embedding for twitter sentiment classification. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.* 2014. pp. 1555–1565.
20. Xiao Z, Liang P. Chinese sentiment analysis using bidirectional LSTM with word embedding. In: *Proceedings of International Conference on Cloud Computing and Security.* 2016. pp. 601–610.
21. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. In: *Proceedings of the 2013 International Conference on Learning Representations.* 2013.
22. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: *Proceedings of Advances in Neural Information Processing Systems.* 2013. pp. 3111–3119. 2013.
23. Packard JL. *The Morphology of Chinese: A linguistic and cognitive approach.* Cambridge University Press; 2000.
24. Chen X, Xu L, Liu Z, Sun M, Luan H. Joint learning of character and word embeddings. In: *Proceedings of the 24th International Joint Conference on Artificial Intelligence.* 2015. pp. 1236–1242.
25. Li Y, Li W, Sun F, Li S. Component-enhanced Chinese character embeddings. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* 2015. pp. 829–834.
26. Yin R, Wang Q, Li P, Li R, Wang B. Multi-granularity Chinese word embedding. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* 2016. pp. 981–986.
27. Yu J, Jian X, Xin H, Song Y. Joint embeddings of Chinese words, characters, and fine-grained subcharacter components. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing.* 2017. pp. 286–291.
28. Su T, Lee H. Learning Chinese word representations from glyphs of characters. In: *Proceedings of 2017 Conference on Empirical Methods in Natural Language Processing.* 2017. pp. 264–273.
29. Cao S, Lu W, Zhou J, Li X. cw2vec: Learning chinese word embeddings with stroke n-gram information. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence.* 2018. pp. 5053–5061.
30. Lass R. *Phonology: An introduction to basic concepts.* Cambridge University Press; 1984.
31. Binyong Y, Felley M. *Chinese romanization: Pronunciation & Orthography.* 1990. Sinolingua Peking.
32. Shi X, Zhai J, Yang X, Xie Z, Liu C. Radical embedding: Delving deeper to Chinese radicals. In: *Proceedings of the 51rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing.* 2015. pp. 594–598.
33. Myers JL, Well A, Lorch RF. *Research design and statistical analysis.* Routledge; 2010.
34. Mikolov T, Yih W, Zweig G. Linguistic regularities in continuous space word representations. In: *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies.* 2013. pp. 746–751. 2013.
35. Levy O, Goldberg Y. Linguistic regularities in sparse and explicit word representations. In: *Proceedings of the 18th Conference on Computational Natural Language Learning.* 2014. pp. 171–180.
36. Khatua A, Khatua A, Cambria E. A tale of two epidemics: Contextual word2vec for classifying twitter streams during outbreaks. *Inf Process Manag.* 2019;56(1):247–57.
37. Fan R, Chang K, Hsieh C, Wang X, Lin C. LIBLINEAR: A library for large linear classification. *J Mach Learn Res.* 2008;9:1871–4.
38. Cambria E, Livingstone A, Hussain A. The hourglass of emotions. *Cognitive Behavioural Systems.* 2012. pp. 144–157.
39. Cambria E, Poria S, Gelbukh A, Thelwall M. Sentiment analysis is a big suitcase. *IEEE Intell Syst.* 2017;32(6):74–80.
40. Poria S, Chaturvedi I, Cambria E, Bisio F. Sentic LDA: Improving on LDA with semantic similarity for aspect-based sentiment analysis. In: *Proceedings of 2016 International Joint Conference on Neural Networks.* 2016. pp. 4465–4473.
41. Susanto Y, Livingstone AG, Chin NB, Cambria E. The hourglass model revisited. *IEEE Intell Syst.* 2020;35(5):96–102.
42. Cambria E, Poria S, Hazarika D, Kwok K. Senticnet 5: Discovering conceptual primitives for sentiment analysis by means of context embeddings. In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence.* 2018. pp. 1795–1802.
43. Çano E, Morisio M. Word embeddings for sentiment analysis: A comprehensive empirical survey. *arXiv preprint* 2019. arXiv:1902.00753
44. Cambria E, Li Y, Xing FZ, Poria S, Kwok K. Senticnet 6: Ensemble application of symbolic and subsymbolic AI for sentiment analysis. In: *Proceedings of the 29th ACM International Conference on Information and Knowledge Management.* 2020. pp. 105–114.
45. Ma Y, Peng H, Cambria E. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence.* 2018. pp. 5876–5883.
46. Ma Y, Peng H, Khan T, Cambria E, Hussain A. Sentic LSTM: A hybrid network for targeted aspect-based sentiment analysis. *Cogn Comput.* 2018;10(4):639–50.
47. Bojanowski P, Grave E, Joulin A, Mikolov T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics.* 2017;5:135–46.

48. Wieting J, Bansal M, Gimpel K, Livescu K. Charagram: Embedding words and sentences via character n-grams. In: Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. 2016. pp. 1504–1515.
49. Goldberg Y, Avraham O. The interplay of semantics and morphology in word embeddings. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. 2017. pp. 422–426.
50. Cao S, Lu W. Improving word embeddings with convolutional feature learning and subword information. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence. 2017. pp. 3144–3151.
51. Cotterell R, Schütze H. Morphological word-embeddings. In: Proceedings of 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2015. pp. 1287–1292.
52. Bhatia P, Guthrie R, Eisenstein J. Morphological priors for probabilistic neural word embeddings. In: Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. 2016. pp. 490–500.
53. Satapathy R, Singh A, Cambria E. Phonsenticnet: A cognitive approach to microtext normalization for concept-level sentiment analysis. In: Proceedings of International Conference on Computational Data and Social Networks. 2019. pp. 177–188.
54. Yang L, Sun M. Improved learning of Chinese word embeddings with semantic knowledge. In: Proceedings of Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. 2015. pp. 15–25.
55. Xu J, Liu J, Zhang L, Li Z, Chen H. Improve Chinese word embeddings by exploiting internal structure. In: Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016. pp. 1041–1050.
56. Peng H, Ma Y, Poria S, Li Y, Cambria E. Phonetic-enriched text representation for chinese sentiment analysis with reinforcement learning. 2019. CoRR abs/1901.07880
57. Huang M, Xie H, Rao Y, Feng J, Wang FL. Sentiment strength detection with a context-dependent lexicon-based convolutional neural network. *Inf Sci.* 2020;520:389–99.
58. Liang W, Xie H, Rao Y, Lau RY, Wang FL. Universal affective model for readers' emotion classification over short texts. *Expert Systems with Applications.* 2018;114:322–33.
59. Valdivia A, Martínez-Cámara E, Chaturvedi I, Luzón MV, Cambria E, Ong YS, Herrera F. What do people think about this monument? Understanding negative reviews via deep learning, clustering and descriptive rules. *J Ambient Intell Humaniz Comput.* 2020;11(1):39–52.
60. Ma Y, Nguyen KL, Xing FZ, Cambria E. A survey on empathetic dialogue systems. *Information Fusion.* 2020;64:50–70.
61. Xu H, Peng H, Xie H, Cambria E, Zhou L, Zheng W. End-to-end latent-variable task-oriented dialogue system with exact log-likelihood optimization. *World Wide Web.* 2019;23(3):1–14.
62. Li X, Xie H, Song Y, Zhu S, Li Q, Wang FL. Does summarization help stock prediction? A news impact analysis. *IEEE Intell Syst.* 2015;30(3):26–34.
63. Picasso A, Merello S, Ma Y, Oneto L, Cambria E. Technical analysis and sentiment embeddings for market trend prediction. *Expert Systems with Applications.* 2019;135:60–70.