Pronunciation-enhanced Chinese Word Embedding

Anonymous submission.

Abstract

Chinese word embeddings have attracted much attention recently. Chinese characters and their sub-character components, which contain rich semantic information, are incorporated to learn Chinese word embeddings. However, Chinese characters are combinations of meaning, structure and pronunciation. Previous works only cover the former two aspects and cannot effectively explore distinct semantics of characters. In this paper, we propose a pronunciation-enhanced Chinese word embedding (PCWE) model, where the pronunciations of context characters and target characters are encoded into embeddings simultaneously. Evaluations on word similarity, word analogy reasoning and text classification validate the effectiveness of our model.

1 Introduction

Word embedding, also known as distributed word representation, represents a word as a real-valued and low-dimensional vector. In recent years, it has attracted great attention and has been applied to many natural language processing (NLP) tasks, such as sentiment classification (Xiong et al., 2018; Yu et al., 2018), question answering (Shen et al., 2017), text analysis (Fast et al., 2017; Le and Lauw, 2017) named entity recognition (E and Xiang, 2017), text segmentation (Zhou et al., 2017; Ma and Hinrichs, 2015) and so on. This benefits from its ability to encode semantic and syntactic information of word into embedding and subsequently makes words with the same or similar meaning being measured closely in a vector space. Among the existing methods, the continuous bag-of-words model (CBOW) and continuous skip-gram model (Skip-gram) are popular because of their simplicity and efficiency, which make it feasible to learn good word embeddings from large corpora (Mikolov et al., 2013a; Mikolov et al., 2013b).

Encouraged from the success of modeling English documents, word embedding has also been introduced to written Chinese text. However, different from English where words are taken as basic semantic units, the characters are usually considered as the smallest meaningful units in Chinese, which are called morphemes in morphology (Packard, 2000). A written character may form a word by itself or on most occasions be a part of a Chinese word. Therefore, Chen et al. (2015) integrated context words with characters to improve Chinese word embedding learning. From the perspective of morphology, a character can be further decomposed into sub-character components, which contain rich semantic or phonological information. Considering the rich internal structural information of characters, many methods have been proposed to improve Chinese word embedding learning by making use of radicals (Li et al., 2015; Yin et al., 2016), sub-word components (Yu et al., 2017), glyph features (Su and Lee, 2017) and strokes (Cao et al., 2017).

The above models improve Chinese word embedding learning from two distinct perspectives of morphology and semantics. Particularly, they explore 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 have different meanings in different words and the semantic information thus cannot be totally revealed from their internal

structures. As shown in Figure 1, "辶" is the radical of "道", but it can merely represent the first meaning. "道" is decomposed into components "辶" and "首", which still semantically related to the first meaning. Also the stroke n-gram feature "首" appears no relevance to any semantics. Although Chen et al. (2015) alleviates this problem by incorporating the character's position information in words and learning position-related character embeddings, this nevertheless cannot reveal distinct meanings. For instance, "道" is used at the beginning of multiple words but with distinct meanings as illustrated in Figure 1.

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

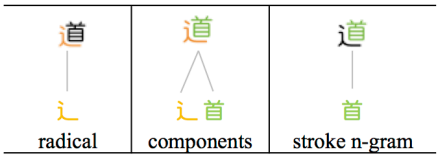

Figure 1: Illustrative example of radical, components and stroke n-gram of character "道".

Chinese is the combination of sound, structure and meaning, corresponding to phonology, morphology and semantics in linguistics, while the above models merely cover the latter two aspects. Phonology describes the way sounds function within a given language to encode meaning[1]. In Chinese, the pronunciation of Chinese words and characters is marked by pinyin[2], which is the official romanization system for Standard Chinese. The Pinyin system includes five tones, i.e., flat tone, rising tone, low tone, falling tone and neutral tone. For example, Figure 2 shows that the syllable "ma" can produce five characters with five tones.

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

Figure 2: Example of five characters with five tones of the syllable *ma*.

However, different from English words where one item only has one pronunciation, many Chinese characters have two or more than two pronunciations, which are called Chinese polyphonic characters. Usually, each pronunciation corresponds to several different but similar meanings. In modern Standard Chinese, one fifth of the 2,400 most common characters have multiple pronunciations[3]. Take a Chinese polyphonic character "长" as an example, as illustrated in Figure 3, it has two pronunciations, i.e., "cháng" and "zhǎng" and each expresses multiple semantics for distinct words. Therefore, we can uncover different meanings of the same character through its pronunciation.

Although a character may represent several meanings for the same pronunciation, we can simply solve this problem by using the other characters' phonological information within the word. This is because the semantic information of the character is determined when collocated

[1] https://en.wikipedia.org/wiki/Phonology
[2] https://en.wikipedia.org/wiki/Pinyin
[3] https://en.wikipedia.org/wiki/Chinese_characters

with other pronunciations. Moreover, the position problem in CWE and MGE is tackled because the position of a character is determined when combined with other characters with current pronunciation. For example, when "长" is combined with "辈 (generation)", they can only form word "长辈" and here "长" represents "old". Although "长" and "年 (year)" could produce two words "年长" and "长年", they can be simply distinguished by the pronunciation of "长", where the former is "zhǎng" and the latter is "cháng" . Hence an intuitive idea is to incorporate pronunciation of characters to learn Chinese word embedding and improve its ability to capture polysemous words.

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)

Figure 3: Chinese polyphonic character "长".

In this paper, we propose a model called pronunciation-enhanced Chinese word embedding (PCWE), which makes full use of information of Chinese characters including phonology, morphology and semantics. The pinyin, which is the phonological transcription of Chinese characters, is combined with Chinese words, characters and sub-character components as context inputs of PCWE. As far as we know, this is the first work that exploits pronunciation of characters for Chinese embedding learning. Evaluation on word similarity, word analogy reasoning and text classification tasks and qualitative analysis are conducted to demonstrate the effectiveness of PCWE.

2 Pronunciation-enhanced Chinese Word Embedding

In this section, we detail our pronunciation-enhanced Chinese word embedding (PCWE), which makes full use of phonological, internal structure and semantical features of Chinese characters based on CBOW (Mikolov et al., 2013b). We do not use Skip-gram model, because CBOW and Skip-gram have few differences and CBOW is slightly faster (Shi et al., 2015). PCWE uses context words, context characters, context sub-characters, and context pronunciation to predict the target word.

We denote the training corpus as D, the vocabulary of words as W, the characters set as C, the sub-characters set as S and the phonological transcriptions set as P. In addition, T represents the size of the context window. As illustrated in Figure 4, PCWE aims to maximize the sum of four log-likelihoods of conditional probability for target word w_i given the average of individual context vectors:

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

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

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

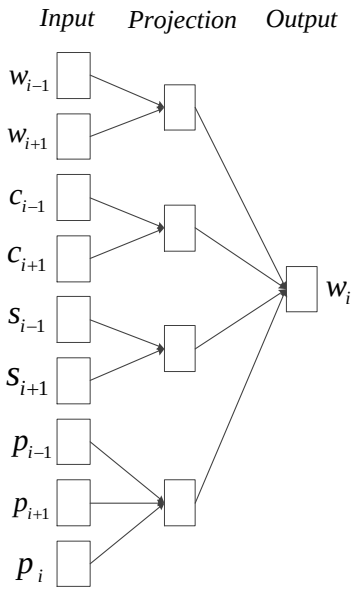

Figure 4: Structure of PCWE, where w_i is the target word, w_{i-1} and w_{i+1} are the context words, and c_{i-1} and c_{i+1} are the context characters. Moreover, s_{i-1} and s_{i+1} are the context sub-characters, p_{i-1} and p_{i+1} denotes the context pronunciation, and p_i is the pronunciation of w_i.

where h_{i1} is the average of context words' vectors, which is estimated as follows:

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

Similarly, h_{i2} is the average of characters' vectors, h_{i3} is the average of sub-characters' vectors, h_{i4} is the average of pronunciations' vector in the context.

There are two differences between JWE (Yu et al., 2017) and our PCWE. First, PCWE integrates the phonological information of words and characters with words, sub-character components to jointly learn Chinese word embedding, while JWE only incorporates words, characters and sub-character components features. In other words, PCWE makes full use of context information from the perspective of both phonology and morphology, while JWE merely uses the morphological features. Second, we utilize the pronunciation of target characters, while JWE incorporates their sub-character components.

3 Experiments

We evaluate our model on word similarity, word analogy reasoning and text classification tasks. For completeness, qualitative case studies are also conducted.

3.1 Datasets and Experiment settings

We employ the Chinese Wikipedia Dump[4] as our training corpus. In preprocessing, we use THULAC[5] for segmentation and pos-tagging. Pure digits and non-Chinese characters are removed. Finally, we obtained a 1GB training corpus with 169,328,817 word tokens and 613,023 unique words. We use the lists of radicals and sub-character components in (Yu et al., 2017), where there are totally 218 radicals, 13,253 components and 20,879 characters.

[4] `https://dumps.wikimedia.org/zhwiki/`

[5] `thulac.thunlp.org/sendMessage_v1_1`

We crawled 411 Chinese pinyin spellings without tones from online Chinese Dictionary website[6]. The Pinyin system includes five tones, and thus we collected 2,055 pinyin spellings in total. We adopted HanLP[7] to transfer Chinese words into Pinyin. Finally, we obtained 3,289,771 Chinese words and pinyin pairs.

We compare PCWE with CBOW (Mikolov et al., 2013b)[8], CWE (Chen et al., 2015)[9], CWE+P, MGE (Yin et al., 2016)[10] and JWE (Yu et al., 2017)[11]. For all models, we used the same parameter settings. We set the context window size to 5, the embedding dimension to 200 and the training iteration to 100. During optimization, we used 10-word negative sampling and fix the initial learning rate to 0.025 and the subsampling parameter to 10^{-4}. Words with frequency less than 5 were ignored during training.

3.2 Word Similarity

This task evaluates the embeddings' ability of capturing the semantic relativity between two embeddings. We adopt two Chinese word similarity datasets, Wordsim-240 and Wordsim-297 provided by Chen et al. (Chen et al., 2015)[12] for evaluation. Both datasets contain Chinese word pairs with human-labeled similarity scores. There are 240 and 297 pairs of Chinese words in Wordsim-240 and Wordsim-297, respectively. However, there are 8 words in Wordsim-240 that did not appear in the training corpus, and 10 words in Wordsim-297 that did not appear in the training corpus. We removed these words and obtained Wordsim-232 and Wordsim-287.

The cosine similarity of two word embeddings is computed to measure the similarity score of the word pairs. We calculated spearman correlation (Myers et al., 2010) between the similarity score computed by word embeddings and the human-labeled similarity score. The higher the spearman correlation is, the better the word embedding in capturing semantic similarity between words. The evaluation results are shown in Table 1.

Table 1: Results on word similarity.

Model	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**

From the results we can observe that PCWE outperforms CBOW, CWE, CWE+P and MGE on the two word similarity datasets. It demonstrates that the combination of morphological, semantical and phonological features can exploit deeper Chinese word semantic and phonological information than other methods. Besides, we further find that PCWE gets competitive results compared to CWE+P. It verifies the benefits of exploiting phonological features to reduce the ambiguity for Chinese character within different words instead of position features. However, PCWE only performs better than all baselines on Wordsim-287, while JWE achieves the best results on Wordsim-232. We think it is probable because the Wordsim-287 contains more Chinese polyphonic characters, such as "行" (háng, xíng) and "中" (zhōng, zhòng).

[6] http://xh.5156edu.com/pinyi.html
[7] https://github.com/hankcs/HanLP/
[8] https://code.google.com/p/word2vec/
[9] https://github.com/Leonard-Xu/CWE
[10] We implement MGE base on the code of CWE.
[11] https://github.com/HKUST-KnowComp/JWE
[12] https://github.com/Leonard-Xu/CWE/tree/master/data

3.3 Word Analogy Reasoning

This task estimates the effectiveness of word embedding to reveal linguistic regularities between word pairs. In this task, given three words a, b and c, the goal is to explore the fourth word d so that a to b is like c to d. Here we use 3CosAdd (Mikolov et al., 2013c) and 3CosMul (Levy and Goldberg, 2014) to find out the nearest word d. We employ the analogy dataset provided by Chen et al. (Chen et al., 2015), which contains 1,127 Chinese word tuples. They are categorized into three types, i.e., capital of countries (677 tuples), state/provinces of cities (175 tuples) and family words (240 tuples). The training corpus covers all words in the analogy dataset. We use accuracy as the evaluation metric and list the results in Table 2 and Table 3.

Table 2: Results on word analogy reasoning measured in 3CosAdd.

Model	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

Table 3: Results on word analogy reasoning measured in 3CosMul.

Model	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

From the results on word analogy reasoning using the 3CosAdd measure function, we can observe that PCWE only achieves the second best performance, while JWE performs the best. Nevertheless, the embedding representations learned by PCWE make better analogy on Capital and City types, when computing by the 3CosMul function. The reason may be that the words in these categories rarely consist of Chinese polyphonic characters, and the morphological features provide enough semantical information for deducing similar word pairs. Furthermore, the results of PCWE are better than those of CWE and CWE+P, which demonstrates the benefit of leveraging compositional internal structure and phonological information.

3.4 Text Classification

Text classification is a common task to evaluate the effectiveness of word embeddings on NLP tasks. We adopt the *Fudan* dataset, which contains documents in 20 topics for training[13] and testing[14]. Following (Cao et al., 2017), we select 12,545 (6,424 for training and 6,121 for testing) documents from 5 topics: environment, agriculture, economy, politics and sports. We average embeddings of the words occurring in documents as the features of documents. We train a classifier by LIBLINEAR[15] (Fan et al., 2008) and show the accuracy of different models in Table 4.

The results indicate that all models achieve over 94% accuracies and our model performs the best. This is because the distinct semantics of characters with different pronunciations are

[13] `http://download.csdn.net/download/github_36326955/9747927`
[14] `http://download.csdn.net/download/github_36326955/9747929`
[15] https://github.com/cjlin1/liblinear

Table 4: Results on text classification in accuracy.

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

captured by our model. For example, for documents in the topic of economy, the word "银行 (bank)" with polyphonic character "行" is used frequently and its pronunciation contributes more than its subcomponents. Therefore, PCWE outperforms other baselines.

3.5 Case Study

In addition to validate the benefit of pronunciation of Chinese characters to improve word embedding, we conduct qualitative analysis by performing some case studies to present the most similar words to certain target words.

Figure 5 illustrates top-10 similar words to two target words for each model. The first example of the target word is "强壮 (qiáng zhuàng, strong)", which includes a Chinese polyphonic character "强" and is used to describe strong man or strong power. The majority of top-ranked words that CBOW generates have no relevance to the target word, such as "鳍状肢 (flipper)" and "尾巴 (tail)". It results from CBOW only incorporates context information, where these words are modified by the target word. Compared to CBOW, CWE exploits many words related to the target word, which contain character "强" or "壮". This verifies the idea of CWE to jointly learn embeddings of words and characters. However, CWE generates word "瘦弱 (emaciated)" that represents opposite meaning to the target word. As the worst, almost all of the words discovered by MGE are not semantically related to the target word, such as "主密码 (master password)" and "短尾蝠 (Mystacina tuberculata)". On the contrary, JWE yields words which are almost all correlated with the target word except for "聪明 (smart)". As the best case, the words identified by our model are all highly semantically relevant to the target word. Besides, only our model generates word "大块头 (big man)", which is related to the target word. Overall, our model can capture semantic relevance well, since it combines the comprehensive information of characters from perspectives of morphology, typography and phonology.

The other example is "朝代 (dynasty)" that also contains a Chinese polyphonic character "朝". It refers to a emperor' s reign or a certain emperor of a pedigree. CWE generates 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 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 an identical situation in the first example, the similar words found by MGE appear no correlation to the target word. However, the majority of similar words generated by JWE are semantically correlated with the target word, yet the words "妃嫔 (imperial concubine)" and "史书 (historical records)" present no relevance to the target word. Conversely, the similar words generated by our model are all highly semantic relevant to the target word.

The singular inclusion of context words information of CBOW leads it to generate contextual words instead of semantic related words. The learning processes of CWE, MGE and JWE involves characters, radicals and internal structures make them tend to identify similar words containing the same characters, radicals and internal structure, while may have no relevant to the target word. In contrast, our model considers the combination of phonological features which can exploit overall information.

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

Figure 5: Case study for semantically related words. Given the target word, we list top 10 similar words for each model.

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

Figure 6: Case study for related Chinese characters and pronunciations. Give a Chinese character, we list top 3 related pronunciation and boldface the right pinyin.

In addition to demonstrate the effectiveness of our model to encode phonology information, we further conduct a case study to exploit the relationship between characters and pronunciations. We first consider analyzing the effectiveness of our model in capturing Chinese polyphonic characters, thus we list related pinyin to the given character as shown in Figure 6. From

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)	泅(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)	崮(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)

Figure 7: Case study for related Chinese characters and pronunciations. Give a pinyin, we list top 5 related Chinese characters.

the results we can observe that PCWE identifies correct pronunciations for every character. For the former six characters, which are polyphonic, our model does not find out all possible pronunciations for the last two ones. The reason may be that the pronunciations not being identified are not commonly used. Figure 7 illustrates relevant characters to the given pinyin. The related characters exploited for the first three pinyin all pronounce as the corresponding pinyin. In Chinese, only character "山" reads as "shān" and our model has identified it. Although the other characters do not sound as the target pinyin, they are all semantically related to hill or mountain. The same situation occurs for the target pinyin "shuǐ". The results show that our model can effectively encode polyphonic features of characters and explore semantically relevant characters.

4 Related Work

In recent years, models specifically designed for Chinese, where characters are treated as the basic semantical units, have been studied. Chen et al. (2015) utilized characters to augment Chinese word embedding and jointly learned Chinese characters and word embeddings. Yang and Sun (2015) took semantic knowledge of characters into account when combining them with context words. There also exists models that incorporate internal structure features of characters to enhance Chinese word embedding learning. Yin et al. (2016) proposed multi-granularity embedding (MGE) by extending CWE with radicals of target words. Li et al. (2015) combined the context characters and their respective radical components as inputs to learn character embeddings. Shi et al. (2015) decomposed contextual character sequence into radical sequence to learn radical embedding. Considering that radicals cannot fully uncover the semantics of characters, Yu et al. (2017) proposed the joint learning word embedding model (JWE) to make full use of context words, context characters and context sub-characters. Moreover, Su and Lee (2017) enhanced Chinese word representation by character glyph features, which learned from the bitmaps of characters. Cao et al. (2017) exploited stroke-level information of Chinese to learn Chinese word embedding. Different from the above studies, our method incorporates multiple features of Chinese characters from aspects of morphology, semantics and phonology.

5 Conclusion and Future Work

In this work, we propose a model called PCWE to learn Chinese word embeddings. It incorporates various features of Chinese characters from perspectives of morphology, semantics and phonology. Experimental results on word similarity, word analogy reasoning, text classification and case studies validate the effectiveness of our model. Although PCWE is proposed to learn word representations for Chinese, it can also be introduced to other languages which share a similar pronunciation system, such as Arabic, where each character may even have more than ten pronunciations. In the future, we plan to apply our model to these languages, and explore comprehensive strategies for modeling phonological information.

References

Shaosheng Cao, Wei Lu, Jun Zhou, and Xiaolong Li. 2017. cw2vec: Learning chinese word embeddings with stroke n-grams. In *Proceedings of the 32th AAAI Conference on Artificial Intelligence*. AAAI.

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huan-Bo Luan. 2015. Joint learning of character and word embeddings. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 1236–1242.

Shijia E and Yang Xiang. 2017. Chinese named entity recognition with character-word mixed embedding. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2055–2058.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Ethan Fast, Binbin Chen, and Michael S. Bernstein. 2017. Lexicons on demand: Neural word embeddings for large-scale text analysis. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 4836–4840.

Tuan M. V. Le and Hady Wirawan Lauw. 2017. Semantic visualization for short texts with word embeddings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2074–2080.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180.

Yanran Li, Wenjie Li, Fei Sun, and Sujian Li. 2015. Component-enhanced chinese character embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 829–834.

Jianqiang Ma and Erhard W. Hinrichs. 2015. Accurate linear-time chinese word segmentation via embedding matching. In *Proceedings of the 53rd 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*, pages 1733–1743.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics*, pages 746–751.

Jerome L Myers, Arnold Well, and Robert Frederick Lorch. 2010. *Research design and statistical analysis*. Routledge.

Jerome L. Packard. 2000. *The Morphology of Chinese: A linguistic and cognitive approach*. Cambridge University Press.

Yikang Shen, Wenge Rong, Nan Jiang, Baolin Peng, Jie Tang, and Zhang Xiong. 2017. Word embedding based correlation model for question/answer matching. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pages 3511–3517.

Xinlei Shi, Junjie Zhai, Xudong Yang, Zehua Xie, and Chao Liu. 2015. Radical embedding: Delving deeper to chinese radicals. In *Proceedings of the 53rd 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*, pages 594–598.

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.

Shufeng Xiong, Hailian Lv, Weiting Zhao, and Donghong Ji. 2018. Towards twitter sentiment classification by multi-level sentiment-enriched word embeddings. *Neurocomputing*, 275:2459–2466.

Liner Yang and Maosong Sun. 2015. Improved learning of chinese word embeddings with semantic knowledge. In *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 15–25.

Rongchao Yin, Quan Wang, Peng Li, Rui Li, and Bin Wang. 2016. Multi-granularity chinese word embedding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 981–986.

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.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xue-Jie Zhang. 2018. Refining word embeddings using intensity scores for sentiment analysis. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 26(3):671–681.

Hao Zhou, Zhenting Yu, Yue Zhang, Shujian Huang, Xin-Yu Dai, and Jiajun Chen. 2017. Word-context character embeddings for chinese word segmentation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 760–766.