Cognitive Computation (2025) 17:75
https://doi.org/10.1007/s12559-025-10431-3

RESEARCH

Attention-Enabled Multi-layer Subword Joint Learning for Chinese Word Embedding

Pengpeng Xue[1] · Jing Xiong[2] · Liang Tan[1,3] · Zhongzhu Liu[4] · Kanglong Liu[5]

Received: 20 June 2024 / Accepted: 16 February 2025

Abstract
In recent years, Chinese word embeddings have attracted significant attention in the field of natural language processing (NLP). The complex structures and diverse influences of Chinese characters present distinct challenges for semantic representation. As a result, Chinese word embeddings are primarily investigated in conjunction with characters and their subcomponents. Previous research has demonstrated that word vectors frequently fail to capture the subtle semantics embedded within the complex structure of Chinese characters. Furthermore, they often neglect the varying contributions of subword information to semantics at different levels. To tackle these challenges, we present a weight-based word vector model that takes into account the internal structure of Chinese words at various levels. The model further categorizes the internal structure of Chinese words into six layers of subword information: words, characters, components, pinyin, strokes, and structures. The semantics of Chinese words can be derived by integrating the subword information from various layers. Moreover, the model considers the varying contributions of each subword layer to the semantics of Chinese words. It utilizes an attention mechanism to determine the weights between and within the subword layers, facilitating the comprehensive extraction of word semantics. The word-level subwords act as the attention mechanism query for subwords in other layers to learn semantic bias. Experimental results show that the proposed word vector model achieves enhancements in various evaluation metrics, such as word similarity, word analogy, text categorization, and case studies.

Keywords Chinese word embedding · Semantic analysis · Attention mechanism · Feature substring · Morphological information · Pronunciation information

Introduction

Word embeddings, or distributed word vector representations, are integral to the landscape of natural language processing (NLP), facilitating the mapping of words or phrases onto a continuous vector space. The quality of word vectors directly affects the performance of various natural language processing tasks. Moreover, an improved presentation of word embeddings can contribute to enhanced performance in downstream tasks, encompassing named entity recognition [1, 2], text categorization [3], sentiment analysis [4] and machine translation [5]. Among the existing word vector methods, the continuous bag-of-words (CBOW) model and the skip-gram model have received widespread attention for their simplicity and effectiveness in learning embedded words in large corpora [6, 7]. The CBOW model predicts the target word by leveraging surrounding contextual words, whereas the skip-gram model predicts contextual words given a target word.

Exploring the disparities between Chinese and English language structures has been a focal point of research. Chinese characters manifest in two dimensions, unlike English, which follows a one-dimensional structure. Chinese charac-

✉ Liang Tan
jkxy_tl@sicnu.edu.cn

[1] School of Computer Science, Sichuan Normal University, Chengdu 610101, Sichuan, China

[2] Chongqing Key Laboratory of Public Big Data Security Technology, Chongqing College of Mobile Communication, Chongqing 401420, Chongqing, China

[3] Institute of Computing Technology, Chinese Academy of Science, Beijing 100190, Beijing, China

[4] School of Mathematics and Statistics, Huizhou University, Huizhou 516007, Guangdong, China

[5] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, 999077
Hong Kong SAR, China

Published online: 03 March 2025

ters leverage the spatial dimensions of the page, enabling movement in two directions—up and down. In contrast, English is read from left to right and does not have such complex geometric variations [8]. These disparities imply that models trained in English may exhibit diminished performance within a Chinese context. Due to the significant differences between the two languages, existing word vector models designed for English do not produce satisfactory results when directly applied to Chinese texts. English is composed of a relatively small number of 26 letters, and the formation of words is relatively straightforward. However, Chinese is a logical language where characters are composed of radicals with specific structures, strokes, and even pinyin, each of which carries a rich semantic meaning.

To address these challenges, recent research on Chinese word embeddings has focused on utilizing subword-level information to capture word semantics. Various approaches have been explored to enhance Chinese word embeddings by leveraging subword information, including characters [9], radicals [10], subword components [11], glyph features [12], and strokes and pinyin [13, 14]. By integrating these subword features, neural network models can capture more nuanced semantic intricacies.

A significant difference between Sino-Tibetan languages, represented by Chinese, and Indo-European languages, represented by English, lies in the natural pattern of inter-word separation. In Chinese, words often lack distinct boundary markers, presenting instead as a continuous flow of characters that collectively form the sentence framework. Therefore, before computer processing and analyzing Chinese text, the first step is to perform effective Chinese word separation in order to accurately distinguish and identify the individual lexical units that make up a sentence. Only after the words are separated will they be input into the model as a complete morpheme to train the word vectors, and the model constructs the vocabulary based on these words. Effective word segmentation necessitates external knowledge to identify specialized entities, exemplified by the classic case of “南京市长江大桥”(Nanjing Yangtze River Bridge). The accurate segmentation should be “南京市/长江/大桥”(Nanjing City/Yangtze River/Bridge), rather than “南京/市长/江/大桥”(Nanjing/Mayor/river/bridge) or alternative formats.

Chinese researchers not only delve into word embeddings based on extensive corpora but also posit that the constituent characters of Chinese words harbor substantial intrinsic knowledge. For instance, consider the Chinese term “道路”(road). Its semantics can be learned from the context of the text corpus, and the meanings of the characters “道” and “路” can also be inferred. The character “道” may indicate a broad avenue, while the character “路” may indicate a winding path. his fusion of character combinations imbues the entire word with distinct semantics. Hence, through the exploration of the internal framework of Chinese characters, we can not only extract semantic insights from contextual cues but also derive nuanced and precise semantic nuances by scrutinizing inter-character relationships.

In this article, “components” refer to segments of Chinese characters, comprising strokes that serve as functional units within the character, also referred to as constituent elements or parts. For example, in Fig. 1, the components of“路” are “足,” “夂,” and “口.” Chinese characters are divided into two kinds of structures: compound characters and unique characters, in which the compound characters are mostly composed of several components, while the unique characters are an integral whole, which can be regarded as a component in itself. Some studies differentiate between radicals and components as distinct sub-word layers. However, this paper contends that radicals are merely a specialized subset of com-

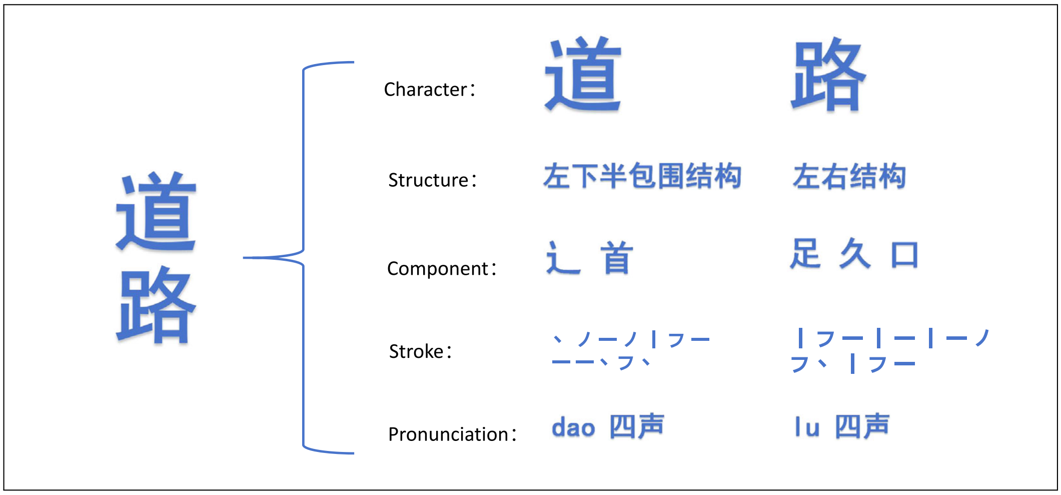

Fig. 1 Diagram of the six-level structure of Chinese words

ponents and should be categorized as such. The components “辶” and “足” mean to walk with one’s feet, which represents part of the semantics of the word “道路.”

Chinese characters exhibit various structural configurations, including top-bottom, left-right, and other orientations. Similar to the distinction between perspective subjects and objects in paintings and primary and secondary strokes in characters, the relative significance of components in different structural positions can also convey semantic nuances within a character. For instance, the character “您” is a honorifics in chinese. The character is a top-bottom structure, in which “你” (you) is positioned above “心” (heart). This structure means “take you to heart” to show respect. Another example is “回” (back), which is an enclosed structure consisting of two “口”(mouth). From the perspective of character interpretation, one meaning is that the two “口” form a loop, and there is always a time to go back to the original point when you go forward. The other meaning is that the outside “口” represents home, and the inside “口” is mouth, meaning to eat, combining to mean to go home for dinner. Both of these interpretations are closely related to the structure of the Chinese character, side by side proving that the structure also comes with the semantics of the character.

Chinese characters are combinations of pronunciation, structure, and meaning, corresponding to phonology, morphology, and semantics in linguistics. In Chinese, the pronunciation of Chinese words and characters is labeled with pinyin, which is derived from the official romanization system of standard Chinese [15]. Unlike English characters, which typically have one pronunciation, many Chinese characters have multiple readings. These characters are known as polyphonic Chinese characters, and each pronunciation can usually refer to several different meanings. For instance, the character “长” has two pronunciations, “cháng” and “zhǎng,” each carrying diverse semantic connotations. Depending on the pronunciation of the same word, different meanings can be found (See Table 1 for details). Some Chinese characters are onomatopoeic, used to imitate the sounds of nature, such as wind, rain, or the cries of animals. For example, when reading the pinyin “hū hū” for “呼呼,” we can infer that it represents the sound of the wind. Moreover, characters sharing similar structures and strokes, such as “土” and “士,” can be differentiated based on their respective pronunciations (“tǔ” and “shì ”). Other components of a Chinese character can be categorized into form-bound and sound-bound, with the form-bound representing the main semantic meaning and the sound-bound character characterizing the pronunciation. For example, consider the character “狗” (gǒu), where the component “犭” represents the concept of canines, and “句” (jù) is sound-bound, with both “ju” and “gou” being possible pronunciations, the latter being applicable in this context. It can be seen that there are linkages between different subwords to better express the semantics.

Table 1 Chinese polyphonic character “行”

Character	Pinyin	Meaning	Example
行	Xíng	Action	步行(walk), 旅行(travel)
		Activity	行为(behavior), 举行(hold)
	Háng	Commerce	银行(bank), 公司(company)
		Trade	行业(industry)

The most granular part of a Chinese character is the stroke, which refers to the various shapes of dots and lines that make up the character and are uninterrupted. The addition of stroke information enriches the study of Chinese word vectors, and the most commonly used method is to extract the stroke sequence of a character, convert it to a numerical sequence, and generate stroke n-gram information using the sliding window method. By utilizing the subword information at the stroke level, the main semantic information of a character can be obtained based on stroke n-grams, such as the most relevant semantic of “您” is “你.”

Many existing models use one or more levels in the six-level structure in addition to words. For example, the CWE model [9], considers characters in the word to contain rich semantics and considers char-level subwords in its model. Meanwhile, the MGE model [10] employs three subwords levels, namely word, char, and component, whereas the PCWE model [14] integrates subwords at the word, char, component, and pinyin levels. However, a few number of models fully exploit the complete six-level structure, potentially overlooking crucial information. On the other hand, existing models often assign uniform significance to every subword level in the Chinese lexicon, neglecting to account for their varying weights. This may introduce noise and affect the generation of word vectors, so it is necessary to use subword information weights to extract key semantics.

To address these challenges, we propose an **A**ttention-enabled multi-layer **S**ubword joint learning **W**ord **E**mbedding (ASWE) model. Our model categorizes Chinese vocabulary into six tiers of subword information: word, char, structure, radical, stroke, and pinyin. It incorporates an attention module that extracts semantics based on the weights assigned to the subword’s intra- and inter-level information. Specifically, we systematically record and code values to subwords such as structure, radicals, and strokes. Strokes are further categorized into n-grams to extract the semantics of specific components in a Chinese character. In pinyin subwords, tones are categorized into five types and assigned corresponding numbers. The inter-layer attention module gauges the similarity between word embedding vectors and elemental embeddings of subwords within the same layer to ascertain a word’s contribution score within that layer. Similarly, the intra-layer attention module computes the similarity between subword embeddings within a specific layer and the elemental embeddings within that layer to derive the intra-layer contribution score.

The main contribution of this paper is to propose a new method for generating Chinese word vectors, namely the Attention-based multilevel Subword joint learning Chinese Word Embedding (ASWE) model. This model classifies Chinese words into six hierarchical levels of subwords: words, characters, components, structures, strokes, and pronunciation. By leveraging the capabilities of the attention mechanism, our model effectively extracts crucial semantic information across and within different layers. A series of experiments covering multiple tasks such as word similarity, analogy, text categorization, and case study have been conducted. The results show that our model not only improves Chinese embedding technology but also adeptly captures Chinese semantics with greater precision compared to the baseline model.

The rest of this paper is structured as follows: in the "Related Work" section, we will briefly introduce the research progress of foreign word embeddings and Chinese word embeddings in recent years; in the "Model" section, we will present our model architecture and the functional structure of each module; in the "Experiments" section, we will evaluate our model in several tasks; in the "Discussion" section, we discuss the necessity and role of Chinese word vector models in the era of large-scale models and comparative analysis; in the "Conclusion" section, we summarize the whole paper and look forward to future work.

Related Work

The theory of word vectors originates from John Rupert Firth's distributional hypothesis that "the meaning of a word can be represented by the distribution of its context." Word vectors, alternatively termed word embeddings, represent a fusion of sophisticated language modeling and feature learning methodologies. Their essence resides in the ingenious transformation of an initially disparate and discrete one-dimensional lexicon of symbols into a denser, continuous multidimensional vector space. In recent years, research techniques for Chinese word vectors have been divided into three main categories: static word vector modeling, dynamic context-dependent word vector generation, and pre-training word vectors for large models.

For static word vectors, Chen et al. [9] first started the research on Chinese word vectors. They highlighted how the semantic essence of Chinese words ties closely to their characters. This led to the development of the character-enhanced word embedding model (CWE), aiming to enrich word representation. At about the same time, Li et al. [16] argued that the radicals of Chinese characters contain a large amount of semantic information, and they concatenated the radicals with the Chinese characters to make predictions, proposing charCBOW and charSkipGram. Xu et al. [17] later noted that CWE's assumption of uniform word contribution overlooks varying semantic impacts, proposing to the SCWE model. Chinese characters can be decomposed into many components, including radicals. R. Yin et al. also recognize that radicals contain rich semantics, but instead of using concat, they add a new layer to CWE, thus proposing the MGE model, which aims at fully integrating information at the word, character, and radical levels. Yu et al. not only use radicals to supplement the semantic representations but also introduce all the components to train the word vectors and propose the JWE model [11]. Delving into the structure of Chinese characters, its most basic constituent unit is the stroke, just as English words are composed of letters. Taking this as a starting point, the Ant Gold Service AI team [18] draws on fasttext's letter n-gram concept and proposes to use stroke n-gram features to train Chinese word vectors. Bing Ma et al. [19] argued that the existing methods ignored the problem of the semantic contribution of the corresponding sub-word units (characters, radicals, and components) to the Chinese words, and then proposed to utilize the attentional mechanism to capture the Chinese semantic structure of words and presented a new framework, the Attention-based Layered Word Embedding (ALWE) model. The main idea of the model is to generate word vectors by utilizing the inter- and intra-layer attention modules to obtain the contributions of subword information and constituents at all levels. Yang et al [14] recently suggested that the existing methods only get semantic information from the internal structure of Chinese characters, which is considered insufficient to capture the semantics. They proposed a pronunciation-enhanced Chinese word embedding learning method, PCWE, based on CBOW, in which the pronunciation of the context character and the target character are simultaneously encoded into the embedding.

Static word vector models have limitations: each word corresponds to only one fixed vector, which makes it difficult to cope with the semantic differences of polysemous words in different contexts, i.e., the static feature restricts the portrayal of lexical semantic flexibility. To address this problem, researchers have proposed the concept of dynamic word vector modeling, aiming to give word vectors the ability to change with context. Among them, the EMLo model [20] is a representative word vector framework. It combines character-level CNN and bidirectional BiLSTM to generate diverse contextualized vector representations of the same word in different contexts, so as to accurately capture and differentiate semantic diversity and context-dependency of words.

For dynamic word vectors, Yang et al [21] proposed a Chinese text sentiment analysis model based on Elmo and RNN.

The model utilizes the Elmo model to learn a pre-trained corpus and then uses recurrent neural networks to extract the deep features of word vectors. Liu et al [22] enhanced Chinese word vectors by GCN (Graph Convolutional Network) and POS (part of speech). GCN is used to construct POS structure maps and extract syntactic structure information of POS features. Meanwhile, multilayer attention is used to distinguish the importance of different features and further update the vector representation of word vectors about the current context.

Amidst the rapid ascent of generative big model technology, a plethora of advanced pre-training word vector methodologies have emerged, encompassing, but not confined to: the MSE model [23], which specializes in entity linking tasks and aims to disambiguate named entities in spoken language; the Generalized Word Vector Model CoROM [24], and the GTE model [25] developed by the Tongyi Lab; Yudao's BCE model [26], renowned for its robust bilingual and cross-linguistic semantic characterization prowess; and the M3E model [27] by the Moka team. Trained on a large corpus, these expansive word vectors not only provide higher accuracy and semantic richness at the word level, but also exhibit robust generalization capabilities, thus demonstrating exceptional performance in countless natural language processing tasks.

Due to the complexity of the ELMo model and the large resources required for training, and the results of static word vector models such as Word2vec are also excellent. Moreover, the word vector study in this paper is to evaluate the impact of multi-layer subword information and attention mechanisms on word representation, so after comprehensive consideration, this paper uses the CBOW of the Word2vec model as the base model for the word vector study.

Model

Chinese language is mainly composed of pictograms and morphophones, and the parsing study of Chinese word meanings is also mainly based on the internal structure and pronunciation of the characters. In this section, we described the details of ASWE, which makes full use of glyph information, internal structure, pronunciation, and semantic features of Chinese characters based on CBOW. The difference between the two approaches of Word2vec is minimal, and Skip-gram is not chosen because CBOW runs slightly faster [28]. ASWE uses contextual words, characters, components, and structures, as well as contextual, pinyin and strokes of the target word to predict the target word.

The proposed ASWE model consists of three main components: the embedding layer, the intra-subword attention layer, and the inter-level attention layer. The embedding layer computes the subword embedding vectors by multiplying the encoded representations of subwords at each layer with their respective embedding matrices. These subwords are encoded from the index in the corresponding subword dictionary. The intra-subword attention layer then applies an attention mechanism within each subword layer to obtain weighted subword representations. This process varies slightly for word-level contexts compared to other subword levels. Specifically, for word-level contexts, self-attention units are employed to adaptively learn context weights, which are then summed to produce a temp target vector. For other sub-word-level contexts, the weight is obtained by the dot product of the sub-word vector and the temporary target vector.

The inter-level attention layer further applies the attention mechanism across the weighted subword vectors to determine the contribution of each subword level to the overall semantic representation of the target word. This process results in the final semantic vector of the target word, which effectively integrates the contributions from all subword levels. Figure 2 illustrates the architecture of the ASWE model.

The ASWE model incorporates both inter-layer and intra-layer attention modules. As illustrated in Fig. 2, these attention mechanisms are of two types: the red attention represents self-attention and the yellow attention represents scaled dot-product attention. The primary difference between them lies in the selection of the query (Q), key (K), and value (V) matrices. In Fig. 3, subplot (a) corresponds to the red attention in Fig. 2, and subplot (b) corresponds to the yellow attention. The self-attention mechanism used in ASWE is fundamentally similar to the one in BERT, as both are based on the scaled dot-product attention mechanism from Transformer models. This approach was chosen for its simplicity and efficiency, allowing the paper to focus on evaluating the effectiveness of these methods rather than analyzing a single method in detail.

Similar to neural network language models, the ASWE model can be trained by optimizing the classification loss. The negative log-likelihood loss function, which serves as the objective function for the ASWE model, is defined in (1). This objective function aims to predict the target word vector w_t using the implicit vector h_t and minimize the negative log-likelihood of the prediction score. Where $\theta = E, E'$ represents the matrix of input and output word vectors.

$$L(\theta) = -log P(w_t|h_t) \tag{1}$$

In our experiment, the negative sampling technique was used to reduce the computational load during the training of the word vector model. When the word list size is large and the computational resources are limited, the training

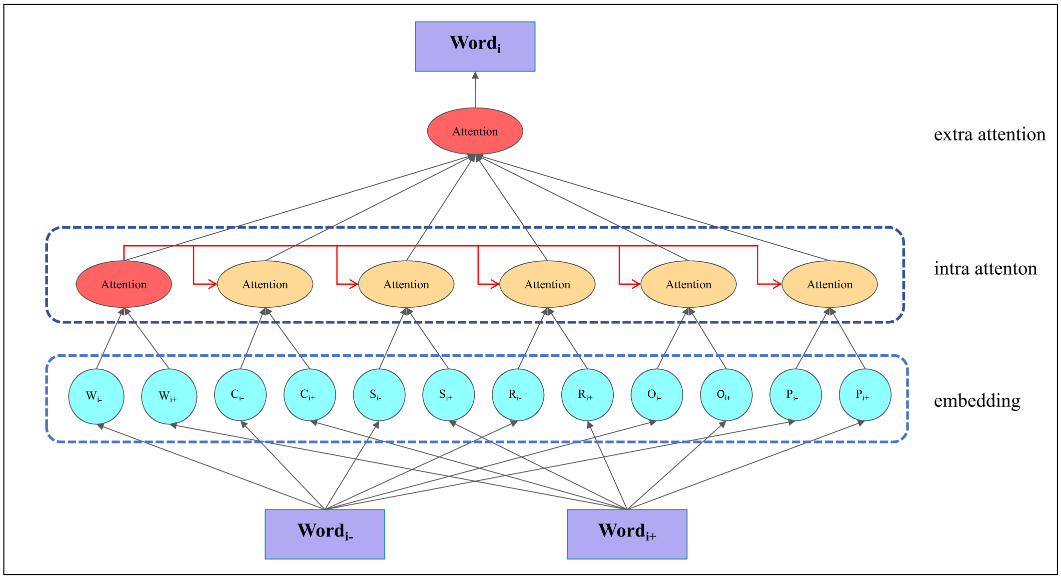

Fig. 2 ASWE model structure

process of general word vector models is affected by the computational efficiency of the output layer probability normalization. The basic idea of the negative sampling approach is to simplify the problem by maximizing the probability of co-occurrence of the two, given the current word and its context. Specifically, we transform the problem into a binary classification problem for the target word and the context. This has the advantage of greatly saving computational resources.

To provide a clearer and more comprehensive understanding of ASWE, its mathematical model is presented next. The following sections will follow the "input-process-output" framework to explain how ASWE utilizes multi-level subword information and attention mechanisms to improve word vector training.

Input

- Corpus D: A large-scale text dataset for model training.
- Target word w: Selected words in the corpus.

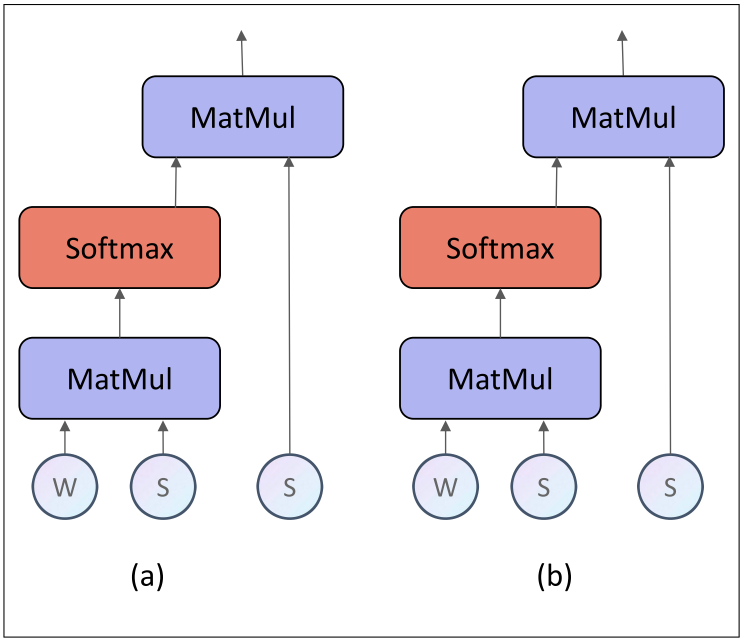

Fig. 3 Attention in ASWE

- The context word set C_w: The context word set of the target word w.
- Embedding Matrices E_s^l for each subword Layer l: Randomly initialized embedding matrices, where each layer l has a specific matrix E_s^l.

Process

(1) Extract the target word and context.
The context words C_w and target words w are obtained from corpus D. The context words C_w are further decomposed into multiple subwords S_C^l across different layers l: where S_C^1 represents word-layer, S_C^2 represents character-layer, and so forth. These subwords are then encoded and batched for further processing.

(2) Negative sampling to obtain negative words.
Use negative sampling to obtain a set of negative words N_w for the target word w.

(3) Compute subword embedding vectors.
For each subword S_C^l compute the embedding vectors x^l using the layer-specific embedding matrix E_s^l.

(4) Intra-subword attention layer.
The famous formula for calculating attention is as (2), this paper uses the attention method in Transformer.

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

The details of this formula are not explained in this article. Interested readers may refer to the original article [29]. The weighted word-level vector uses the self-attention mechanism, i.e., Q, K, and V are the same, so the weighted word-level vector is calculated as:

$$h_t^1 = Attention(x^1, x^1, x^1) \tag{3}$$

The other weighted sub-word vector h_t^i is calculated by the weighted word-level vector to do the attention query:

$$h_t^i = Attention(x^1, x^i, x^i), where\ i\ in[2, 6] \tag{4}$$

(5) Global attention on hierarchical subword vectors.
The final hidden layer vector h_t is also derived from the self-attention mechanism and is calculated in a manner similar to (3) except that it differs from the input $Concat(h_t^1, h_t^2, ..., h_t^6)$. The hidden layer vector h_t is the same as h_t in (1).

(6) Loss calculation
Compute the final vector h_t and multiply it with the target word and its negative samples, applying the log-sigmoid function to compute the loss L:

$$L = -\sum_{w \in D}\left(\log(\sigma(h_t \cdot w)) + \sum_{n \in N_w}\log(\sigma(-h_t \cdot n))\right) \tag{5}$$

Output

- Optimized Subword Embedding Vectors E_s^l: The resulting subword embedding matrices after training, adapted for semantic representation.
- Minimized Loss Value L: The loss value represents the semantic relevance between the target word and its context and negative samples.

This objective function is an extension and supplement of PCWE and ALWE. PCWE does not take into account the use of the attention mechanism to reduce the noise of the subword information at each level, and the subword information of ALWE is not perfect enough. Whereas, ACWE fully absorbed the features of similar excellent models, supplemented structural and stroke subword information, and improved the subword information at each level of Chinese words while using the attention mechanism to minimize the interference of meaningless information, and captured meaningless information at the embedding level of Chinese words. At the same time, ASWE utilized the attention mechanism to reduce the interference of meaningless information as much as possible and captured more accurate semantics.

The model architecture is shown in Fig. 2. The model first obtained the base vectors of the words through the self-attention mechanism as a benchmark for the similarity of words, structures, components, pinyin, and strokes. The subwords in the layer and the benchmark vector were weighted according to the cosine similarity, and the vectors of the subwords in the layer were obtained according to the weights. Finally, all the subword vectors and word vectors were used to get the final attention word vector using the attention mechanism.

Experiments

Corpus and Parameter

To build the training corpus, we utilized Chinese Wikipedia dumps,[1] employing scripts from the gensim toolkit[2] for dataset conversion during data preprocessing. Due to the

[1] https://dumps.wikimedia.org/zhwiki/

[2] https://github.com/piskvorky/gensim

presence of numerous traditional Chinese characters, we applied the OpenCC[3] toolkit to simplify them. Jieba[4] was chosen for word segmentation due to its widespread usage, user-friendly API, and high efficiency. Eliminating non-Chinese characters and numbers, we filtered out sentences with fewer than 5 words. Ultimately, we acquired 233,666,330 lexical tokens and 2,036,032 unique words, amounting to a training corpus of 1.67G. Additionally, we referenced the Han Dian website https://www.zdic.net/ compilation of components and strokes, which includes 13,252 components, 5 strokes, and 20,940 characters.

The structural dataset[5] is the sole open-source resource discovered for this purpose. Nevertheless, the dataset incorrectly classifies structures into 14 categories, actually there being only 13 distinct classes. Consequently, this leads to redundant classifications, exemplified by the semi-envelope structure, which should be further divided into upper left semi-envelope, lower left half-envelope, and several other specific structures. To rectify these inaccuracies, we manually corrected 534 errors within the dataset.

We conducted word-to-pinyin mapping, incorporating pinyin[6] atop pcwe to supplement words lacking pinyin annotations. Chinese pinyin devoid of tones was assigned the neutral tone "5." This process yielded a total of 3,539,391 pinyin pairs for Chinese words.

ASWE is evaluated against analogous models PCWE, CWE, MGE, and JWE, employing identical parameter configurations for equitable comparison. Parameters were set as follows: context window size of 5, word vector dimension of 200, 100 iterations, 10 negative samplings, initial learning rate of 0.025, minimum learning rate of 0.0001, and a subsampling rate of 1e-4.

Experimental Details

Strokes are the most fine-grained units in Chinese characters, and each Chinese character can be decomposed into a sequence of strokes with a specific order. According to the inspiration of [30], we categorize strokes into five types: horizontal, vertical, apostrophe, dot, and fold, which are also the five basic types of strokes stipulated in the "Common Character List of Modern Chinese" jointly issued by several Chinese cultural departments in 1988. We coded these 5 types of strokes as shown in Table 2.

Table 2 Stroke code

Strokes	Horizontal	Vertical	Left-falling	Right-falling	Turning
Shape	一	丨	ノ	丶	フ
Code	1	2	3	4	5

Chinese characters' structure comprises strokes and orientation relationships among components, with 13 structure types coded as depicted in Table 3. Morphologically, Chinese characters fall into two categories: monograms and composite characters. Monograms consist of a single constituent element with an indivisible structure, while composite characters amalgamate two or more components. Occasionally, characters with distinct structural features are labeled differently in academia. For instance, characters like "品," "晶," "森," exhibit balanced upper, middle, and lower parts and are metaphorically termed "Pin zigzag structure" by certain scholars. However, this paper adopts a broader classification criterion of upper and lower structure to systematically explore and analyze Chinese character structures.

The pinyin system encompasses five fundamental tone classifications: the flat tone, noted for its smooth and calming pitch; the rising tone, characterized by an upward inflection; the falling-rising tone, marked by a fluctuation between descending and ascending tones; the falling tone, distinguished by a pronounced descent in syllabic pitch; and the neutral tone, softly pronounced without a fixed tonal value. Tonal symbols are typically positioned above or after the main vowel of the syllable for labeling purposes, with lighter tones notably lacking specific labels. Tones are coded numerically from 1 to 5, as illustrated in Table 4.

Word Similarity

The measure of word sense relevance is one of the important properties of word vectors. Word vectors can be measured according to their ability to express word sense relevance.

Table 3 Structure of Chinese characters

Structure	Shape	Structure	Shape
Left-right	林	Left surrounding	区
Left-middle-right	慨	Left-upper surrounding	友
Up-down	客	Left-bottom surrounding	这
Up-middle-down	意	Right-upper surrounding	匈
Entire surrounding	囚	Mosaic	夷
Upper surrounding	冈	Integral	女
Bottom surrounding	函		

[3] https://github.com/BYVoid/OpenCC

[4] https://github.com/fxsjy/jieba

[5] https://github.com/yunzhangwww/feature-database

[6] https://github.com/hotoo/pinyin

Table 4 Pinyin example

Tone	Flat	Rising	Falling-rising	Falling	Neutral
Example	mā	má	mǎ	mà	ma
Character	妈	麻	马	骂	吗

The relevance between any two words can be easily measured by utilizing the low-dimensional, dense, and continuous properties of word vectors. For example, given words w_a and w_b, their cosine similarity within the word vector control can be used as a measure of their lexical relevance:

$$sim(w_a, w_b) = \cos(v_{w_a}, v_{w_b}) = \frac{v_{w_a} \cdot v_{w_b}}{\|v_{w_a}\| \|v_{w_b}\|} \tag{6}$$

The word similarity task aims to evaluate the performance of word vectors in capturing semantic relatedness. We use two Chinese word similarity datasets, Word-sim240 and Word-sim297, provided by Chen et al [9] for evaluation. These two datasets contain a series of Chinese word pairs, each with manually labeled similarity scores. There are 240 and 297 Chinese word pairs in Wordsim-240 and Wordsim-297, respectively. The similarity between two-word embeddings is measured by calculating the cosine similarity between them. This method can effectively measure the angle between two-word vectors, thus reflecting their proximity in the semantic space. We calculated the Spearman correlation between similarity scores using word embeddings and manual labeling [31]. Higher Spearman correlation values indicate that word embeddings capture the semantic similarity between words more effectively. By analyzing the Spearman correlation, we can determine whether the word embedding model can accurately reflect human perception of semantic similarity. The evaluation results are shown in Table 5.

CWE+P in the table represents the addition of positional coding to the CWE model. Some of the model experimental results used are data from the corresponding papers, even so, the effect of our model is quite good after comparison.ASWE i.e., our proposed model based on six layers of subwords and full attention. To validate the effectiveness of the model, we conducted ablation experiments. In particular, ASWE-AT represents the removal of the inter- and intra-layer attention modules relative to ASWE, using summation as well as averaging instead; ASWE-S represents the ablation of the two added layers of stroke and structure subwords (compared to PCWE), but retaining the attention module. These experiments aim to explore the effects of different components on model performance. The visualization results are shown in Fig. 4.

Table 5 Word similarity results

Model	Wordsim-240	Wordsim-297
CBOW [7]	0.5322	0.5746
CWE [9]	0.5138	0.6022
CWE+P	0.5075	0.5960
MGE [10]	0.4635	0.5231
JWE [11]	0.5246	0.5641
ALWE [19]	0.5487	0.5628
PCWE [14]	**0.5542**	0.6087
ASWE-AT	0.5329	0.6140
ASWE-S	0.5474	**0.6342**
ASWE	0.5434	0.6254

The experimental results show that ASWE outperforms most models on both datasets. This indicates that using a combination of deeper morphological, semantic, and phonological features provides better access to the semantics of words. Specifically, ASWE and its ablative variant experiments performed slightly worse than PCWE on the Wordsim-240 dataset but obtained the best review results on Wordsim-297. And among the three different forms of ASWE, the best results were obtained by ASWE-S. ASWE-S, i.e., the four layers of words, characters, components, and pinyin subwords, introduces the inter-layer attention module and the intra-layer attention module to learn a weight-based representation of contextually relevant word vectors and obtains a semantic distribution of words in a context by automatically adjusting the relationship between the words in the context and other words so that words that are related to the context have the importance of words with different semantic relevance is weighted differently. In addition, the importance of subwords in each layer is weighted according to their similarity to the word vector to avoid semantic interference and noise caused by subword information that does not contribute much to the semantics of the word.

Word Analogy

Word analogy is another common internal task evaluation method for word vectors. Analyzing the distribution of word vectors in the vector space, it can be found that suppose we have two-word pairs (w_a, w_b) and (w_c, w_d), which have the same relationship syntactically or semantically. In other words, these two-word pairs have similar meanings or functions in a particular context. Based on the properties of word vectors, we can observe the geometric properties of $v_{w_b} - v_{w_a} \approx v_{w_d} - v_{w_c}$. Example:

$$v_{king} - v_{man} \approx v_{queue} - v_{women} \tag{7}$$

This equation shows that for two-word pairs (king, man), (queen, woman) with the same relation, there is the same logical relation between their semantics, so the word vectors

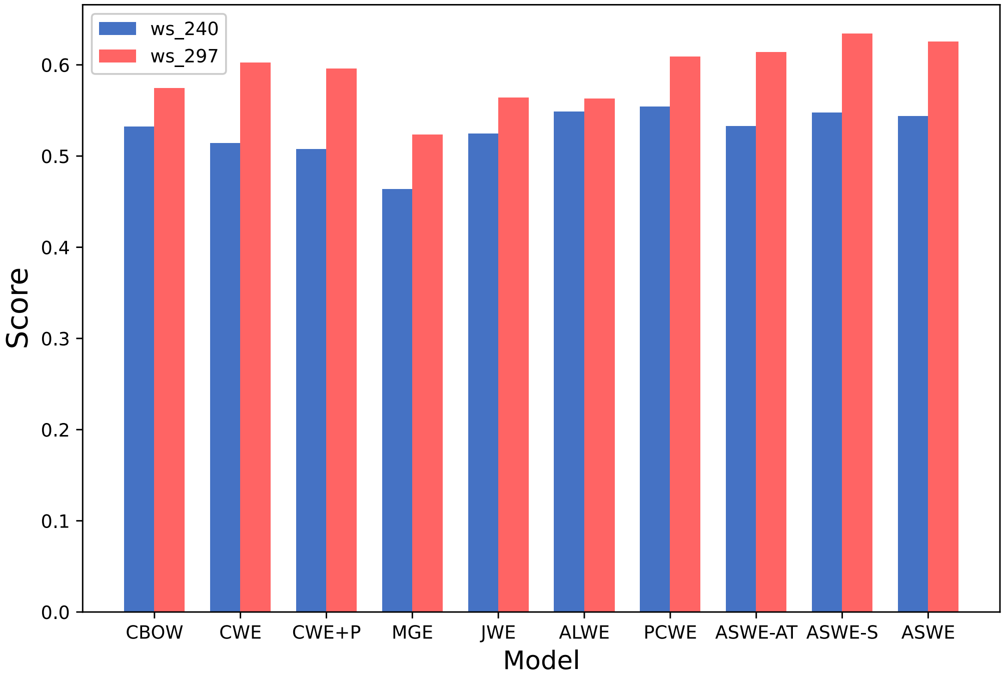

Fig. 4 Results of word similarity task

are relatively close to each other in the vector space. If we subtract the word vectors of the first word pair, the result obtained is very close to the result obtained by subtracting the word vectors of the second-word pair.

We use the analogical reasoning dataset constructed by Chen et al. in CWE, which contains 1124 sets of evaluation datasets for Chinese analogical questions, divided into 3 themes: Capitals (677 sets), State (175 sets), and Family (272 sets). The final evaluation index is the correctness of reasoning. The experimental results are shown in Table 6, where the symbols of each model are consistent with Table 5.

Throughout this study, we meticulously scrutinized the performance of every model in a word analogy task. Initially, our ASWE and its diverse iterations continue to excel in this task, surpassing the majority of competing models, indicative of their superior efficiency and accuracy in handling such tasks. Additionally, it is worth mentioning that although PCWE exhibits superior performance on the Wordsim-240 dataset, it lags behind ASWE in terms of semantic inference capabilities. This disparity may stem from PCWE's optimization strategy, which, despite targeting specific aspects, falls short of comprehensively enhancing performance across all tasks. While the models exhibit comparable performance on capital and city data, their performance significantly deteriorates on household data compared to the former two datasets. This implies that current models still lack proficiency in character relationship reasoning. Furthermore, results from ablation experiments indicate that ASWE outshines its ablated counterparts, ASWE-AT and ASWE-S. Notably, ASWE-AT, focusing solely on subword information, closely resembles the original ASWE in performance. Conversely, the ASWE-S model, excelling in word similarity evaluation, exhibits relatively inferior performance in word analogy tasks, falling short even compared to the PCWE model. This raises a question that deserves deeper investigation. Maybe ASWE-S's prowess lies in its adept handling of word similarity tasks via its accentuated attentional mechanism, effectively capturing contextual semantic information. Nevertheless, in word similarity tasks, deeper subword information seems pivotal for enhancing the generalizability of word vectors. Consequently, ASWE-S might exhibit slight inadequacy in generalization owing to its emphasis on contextual semantics, performing relatively inferiorly compared

Table 6 Word analogy results

Model	Total	Captical	City	Family
CBOW	0.6699	0.7622	0.7200	0.4081
CWE	0.7687	0.8744	0.8800	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
ALWE	0.6200	0.6380	0.8460	0.4300
PCWE	0.8176	0.8907	0.9029	0.5809
ASWE-AT	0.8345	0.9217	0.92	**0.8345**
ASWE-S	0.8132	0.8892	**0.9257**	0.5514
ASWE	**0.8407**	**0.9291**	0.92	0.5699

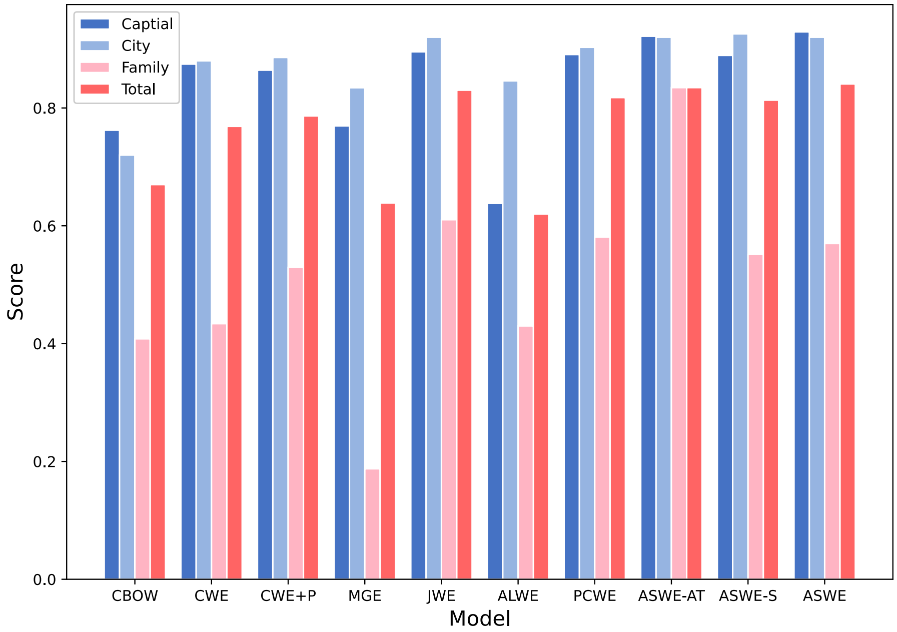

Fig. 5 Results of word analogy task

to ASWE and ASWE-AT, which incorporate multiple layers of subword information. This underscores the importance of considering task-specific characteristics and balancing the relevance of different information in enhancing overall performance when designing word vector models. The visualization results are shown in Fig. 5.

Text Classification

Text categorization is widely used to evaluate the effectiveness of word embeddings in NLP tasks [32]. We use the Fudan Chinese text dataset,[7] which contains documents on 20 topics, for training and testing. After [18], we select 12,545 (6424 for training and 6121 for testing) documents in five topics, namely, environment, agriculture, economy, politics, and sports. We replaced the trained word vectors with word vectors of the same words in the dataset, and the others with randomly generated word vectors, and then froze the gradient update of the embedding layer. For simplicity, we trained TextCNN as a classifier, and the models used all achieved the best classification results in about 30 rounds, as shown in the table below. As shown in Table 7, all the methods achieved more than 98% accuracy, and our method performed the best. This is because our method not only captures the semantics with different subword structures, but also analyzes the weights of the subword hierarchies in them, and ACWE outperforms other baselines.

Case Study

In this paper, we not only validate the value of Chinese character subword information in improving the quality of word embedding expressions but also provide an in-depth comparison of the efficacy of different approaches in identifying words that are closest to the meaning of a particular target word through a series of meticulous case studies. In Table 8, we detail the top 10 most similar lexical examples identified by each model for the two target words.

Taking “强壮” as an example, the word contains “强,” which vividly characterizes an individual as physically fit or powerful. Although the CBOW model relies heavily on contextual information to construct word embeddings, the top-ranked words identified by the CBOW model, such as “前肢” and “尾巴,” are not semantically closely related to “强壮.” The results generated by CWE, on the other hand, contain more character elements of “强” and “壮,” which confirms the effectiveness of the idea of combining word and character embedding in CWE to a certain extent; however, it is worth noting that the CWE also contains words such as “瘦弱,” which are contrary to the original meaning. In

Table 7 Accuracy of text classification

Model	Accuracy
PCWE	98.50%
ASWE-AT	98.66%
ASWE-S	98.59%
ASWE	98.63%

[7] https://download.csdn.net/download/weixin_42691585/12751311

Table 8 Case study for semantically related words

word	CBOW	CWE	MGE	JWE	PCWE	ASWE
强壮	健壮	健壮	主密码	健壮	强健	健壮
	结实	粗壮	强健	强健	健壮	高大
	强健	强健	金眸	粗壮	粗壮	强健
	鳍状肢	身强体壮	短尾蝠	高大	结实	瘦弱
	尾巴	壮健	健硕	结实	壮健	魁梧
	矮胖	壮硕	锐利	体格	壮健	强大
	前肢	坚韧	体格	壮硕	壮硕	瘦削
	短尾巴	肥壮	马羚亚科	聪明	敏捷	瘦削
	身躯	瘦弱	波塞东龙	壮健	吃苦耐劳	吃苦耐劳
	长尾巴	凶猛	腕力	勇猛	白皙	肌肉发达
朝代	历朝	历朝历代	藩属国	封建王朝	历朝	文景之治
	各朝	历代	历朝	王朝	封建王朝	总录
	隋唐	历朝	萍踪侠影录	历朝	王朝	各朝
	分封制	两朝	类书	历朝历代	各朝	大燕
	前朝	诸侯国	年号	嫔妃	历朝历代	王朝
	历朝历代	大统历	盛衰	属明	年号	封建王朝
	典章制度	封建王朝	封建王朝	各朝	丁朝	历朝
	历代	列朝	叶榆县	两朝	吴朝	夷狄
	明清	各朝	相权	历代	历代	国祚
	封建王朝	统元历	嫡庶	君主	大燕	建都

Given the target word, the top 10 simlar words identified by each method are listed

contrast, the MGE method performs relatively poorly, and its generated words such as “主密码” and “短尾蝠” are less semantically related to “强壮,” while the JWE shows higher accuracy. JWE, on the other hand, shows higher accuracy and most of the words have strong semantic associations with “强壮,” except for some words such as “聪明,” and “敏捷” generated by PCWE, “吃苦耐劳” and “白皙” do not match the context of “强壮.” However, our ASWE model has improved compared to PCWE, although there are still incomplete matches like “吃苦耐劳,” the generated words are more relevant to the target words overall.

Taking “朝代” as an example, the polyphonic character “朝” in this word represents the historical dynasty or the era ruled by a certain emperor. In this context, although the CBOW algorithm identifies high-frequency words such as “分封制” and “典章制度” that appear in the textual environment together with “朝代,” it is not satisfactory in terms of semantic accuracy and correspondence. However, the semantic accuracy is not satisfactory. Similarly, CWE is also limited in that the terms it found related to the calendar theme, “统元历” and “大统历,” do not accurately reflect the meaning of “朝代.” Like the result for “强壮,” MGE fails to effectively capture the words closely related to their semantic meaning when analyzing “朝代.” In the results of JWE, except for some words such as “属明,” most of the words are clearly semantically related to “朝代.” As for PCWE, it incorrectly generates “丁朝” and “吴朝,” which did not exist in history. After ASWE excluded the irrelevant term “General Records,” the rest of the words generated by ASWE better reflected the semantic association with the concept of “朝代.”

In summary, the CBOW model relies solely on contextual word information, which in some cases may lead to the generation of words with weak semantic associations with the target words. CWE, MGE, and JWE take into account the character, radical, and internal structure information in the training process, but this may lead to the limitation of retrieving words with similar meanings to the target words based on the sharing of the same characters or structural features while ignoring the actual differences in the meanings of the words. PCWE tries to enhance the level of information about the characters by integrating their phonological attributes, whereas our proposed ASWE model outperforms the above-mentioned model in the overall performance of case study analysis. The overall performance of our proposed ASWE model outperforms the above models in case analysis, but it still needs to focus on and improve the problem of semantic fit between individual generated words and target words.

In evaluating the effectiveness of word vectors in characterizing semantic relevance, we employed a systematic methodology. To aid visualization, we randomly selected 50 common nouns for analysis. Utilizing the PCA technique, we projected these vectors onto a two-dimensional space for visual representation. In Fig. 6, closer proximity between words indicates stronger semantic relevance. The illustra-

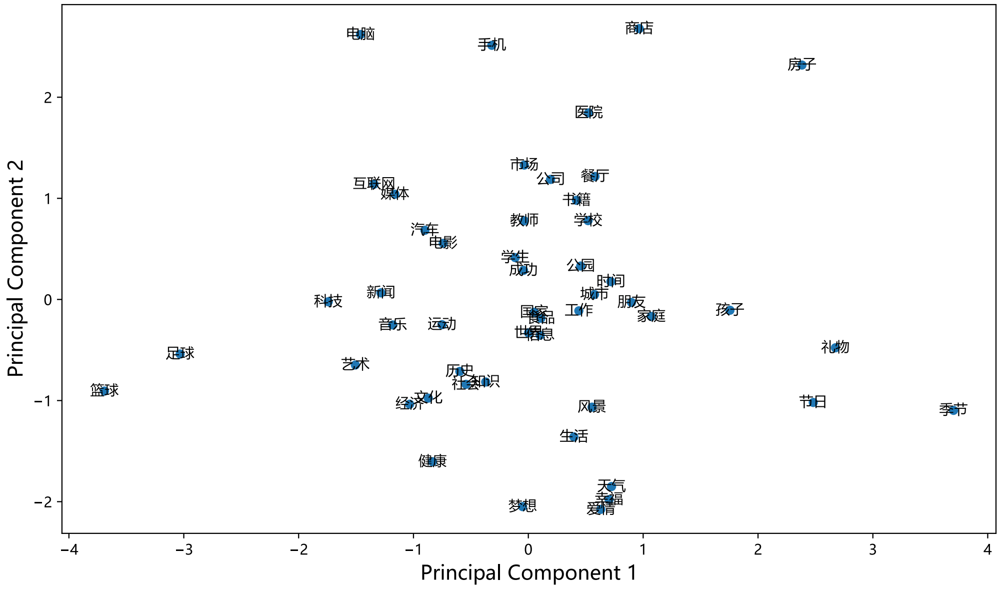

Fig. 6 2D Projection of Word Vectors via PCA

tion unveils notable semantic clustering. For instance, the closer proximity between “国家”(country), “世界”(world) and “信息”(information) indicates their high semantic relatedness, aligning with human cognitive and linguistic patterns. Likewise, “历史”(history), “社会”(society) and “知识”(knowledge) exhibit significant clustering, affirming their semantic relatedness. This suggests that ASWE word vectors capture both surface and structural semantic relationships.

Complexity and Time Consumption

Our model ASWE adds stroke and structure sub-word information and uses intra-and inter-layer attention mechanisms to enhance Chinese word vector training. These measures will result in additional time overhead and memory stress. In view of scientific rigor, we choose PCWE, the main reference model of this paper, to compare the algorithm complexity and training time-consuming.

The structure of the PCWE model is embedding layer, average layer, and output layer. The input of the embedding layer includes words, characters, components, and pinyin. The average layer computes the average of these subwords to get the intermediate vector, and the output layer computes the similarity of the intermediate vector to the target word and the negative sample word and then computes the log softmax loss.

Analyzing the dataset for this paper, each word contains an average of 3.22 characters and 1 pinyin, and each Chinese character contains an average of 1.79 components, 1 structure, and 12.84 strokes. For the following analysis, we set the word vector dimension to 200, the context window to $2 * 5 = 10$, and the number of negatively sampled words to 10.

On memory pressure, because of the large size of corpora, both PCWE and ASWE models show the phenomenon that the parameters of the embedding layer are much larger than those of other layers, this results in a small difference between the two models in the learnable parameters. The difference in memory pressure between the two models is mainly due to input, as ASWE’s stroke layer grows with word size. The input size of PCWE was calculated to be approximately $209.84 tensor$, whereas the input size of PCWE was approximately $655.49 tensor$, 212.38% larger than the former.

In computation, the embedding layer is mainly an index lookup operation, without a large number of multiplication or addition operations, while the output layer is the same two models, so the difference of computational load mainly lies in the middle layer, that is, the average layer of PCWE and the attention layer of ASWE. After calculation, PCWE had a computational load of approximately $65M$ floating-point calculations per word, while ASWE had a computational load of $90M$, which was 38.46% larger than the former. We then re-ran the PCWE and ASWE models separately and recorded their training duration, with PCWE taking 257 min, and ASWE taking 368 min, 43.2% more than the former. It turns out that the training time is longer than it should be because PCWE uses more sub-word information and has a larger input dimension, leading to more frequent memory access during model training. A comparison of complexity and time is shown in Fig. 7.

Although our model ASWE has increased the algorithm complexity and training time compared to the baseline

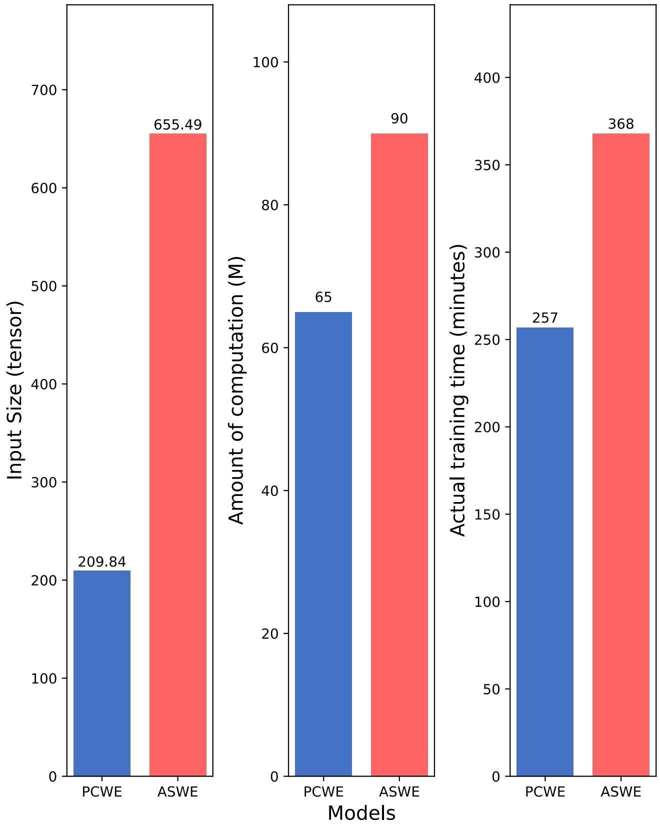

Fig. 7 Complexity and time comparison

model, it also improves the performance of our model in many tasks, such as word similarity, word analogy, and text classification. Performance and complexity cannot be both, and we think the increase in complexity is worth it. Of course, we will also consider optimizing the time performance of the model in the future.

Discussion

The Necessity of ASWE in the Era of Large-Scale Models

Large pre-trained models such as BERT and GPT have demonstrated outstanding performance in various natural language processing tasks, thanks to their ability to utilize rich contextual information for semantic inference. In Chinese language processing, some challenges such as polysemy, synonymy, and idiomatic expressions often necessitate the inclusion of structural information from individual characters, the embedding of Chinese characters remains crucial and has unique advantages in the following aspects:

- Problem of polysemy: There are a large number of polysemous words in Chinese, and their meanings often depend on the internal structure and context of the characters. Large pre-trained models typically infer the meaning of polysemous words through context, but in some specific contexts (such as short text length, unclear context, etc.), they may not accurately capture subtle semantic differences. Chinese character embedding can provide richer semantic information by refining the character layer and its constituent elements (such as pinyin, strokes, etc.), thereby helping to overcome some contextual limitations in handling polysemous words.
- Idioms and fixed collocations: Chinese idioms, idioms, and other fixed collocations are often determined by the combination of words and grammatical structures. Although large-scale models can capture these combinations in certain situations, they may be more difficult to handle Chinese idioms, classical texts, and so on. By modeling the hierarchical structure of characters, ASWE can provide more accurate semantic representations for these complex language phenomena, thereby enhancing the unique advantages of the model in Chinese processing.
- Complexity of language phenomena: The language phenomena in Chinese are rich and diverse, such as the construction of semantic words, antonyms, and character shapes, which pose higher requirements for word embedding models. ASWE can effectively improve the accuracy of semantic representation by capturing this information at different levels, such as pinyin, drawing, etc.

Comparative Analysis with Existing Large-Scale Pre-trained Models

Although large pre-trained models such as BERT and GPT have shown excellent performance in various natural language processing tasks, they still have some limitations:

- Context dependency problem: Current large-scale language models (such as BERT, GPT, etc.) rely heavily on contextual information to infer word meanings in Chinese processing, especially in understanding phenomena such as polysemy, synonyms, idioms, etc., which often require the combination of character structure information. The ASWE model refines semantic expression by embedding words at different levels (such as pinyin, radicals, etc.), which can extract key information from the internal structure of words even in the absence of context.
- Lack of generalization ability: Although large models perform well in various tasks, their generalization ability may be limited in certain specific tasks. For example, when dealing with more specialized text types such as classical Chinese, poetry, literary works, etc., pre-trained models may not fully understand the deep meaning of the text, while ASWE can capture these complex phenomena through the hierarchical structure of characters and provide more refined semantic representations.

Comparative ASWE model with existing large-scale pre-trained models, such as BERT, GPT, etc. The advantages of the ASWE model are as follows:

- Advantages of Short Text Semantic Representation: The ASWE model has significant advantages in the semantic representation of short text, especially in certain tasks such as sentiment analysis, keyword extraction, and short text classification. It effectively avoids the interference of contextual information in long texts on the understanding of short texts and improves the accuracy of semantic understanding in short texts by weighting information at the level of sub words.
- Performance of specific tasks: For example, in text classification tasks, ASWE can accurately capture detailed information through refined sub-word structures, providing more accurate results than traditional large-scale models. Especially in tasks that require fine-grained semantics.

All in all, although this study does not claim that ASWE outperforms large pre-trained models in word embedding, it proposes a promising approach that, when integrated with these models, could enhance their performance. The primary contribution of this paper is to demonstrate the potential of the six-layer semantic structure and attention mechanisms in enriching word embedding. Future research will explore how incorporating these elements into large models can further improve results.

Conclusion

In this research, we introduced a novel approach named ASWE for learning Chinese word embeddings. ASWE integrates various features of Chinese characters, including characters, components, strokes, structures, and pinyin, from morphological, semantic, and phonological perspectives. It employs multiple attention mechanisms to determine the weights among these features, resulting in a weighted-word vector representation. Ultimately, the effectiveness of ASWE is validated through extensive experiments, including word similarity, word analogical reasoning, text categorization, sentiment analysis, and case studies. The experimental findings reveal that incorporating additional Chinese character subwords enhances the analogical capabilities of Chinese word vectors, whereas employing attention mechanisms facilitates the acquisition of more hierarchical semantics. For future research, We intend to extend the concept of ASWE to dynamic word embeddings and large-scale pre-trained word embedding algorithms. And the concept of sub-word is applied to the semantic representation of the large model, which will help the large model to analyze and deal with short texts, complex ancient Chinese characters, poetry, and other scenes more effectively.

Acknowledgements Dear editor, thank you for processing the manuscript!

Author Contributions Conceptualization, data curation, methodology, software, writing — original draft by Pengpeng Xue; formal analysis, investigation by Jing Xiong; funding acquisition, supervision, project administration, writing — review and editing by Liang Tan; supervision, supervision, validation, writing — review and editing by Kanglong Liu; resources, supervision by Zhongzhu Liu, All authors have read and agreed to the published version of the manuscript.

Funding This work was supported by the Sichuan Provincial Science and Technology Department Project under Grant (No. 2022YFG0161 and No. 2023YFG0295). This research was also supported by the National Natural Science Foundation of China under Grant (No.613731 26). We would also like to thank the anonymous reviewers for their constructive and insightful feedback.

Data Availability No datasets were generated or analyzed during the current study.

Declarations

Conflict of Interest The authors declare no competing interests.

Generative AI and AI-Assisted Technologies in the Writing Process Statement: During the preparation of this work the author(s) used Chat-

GPT in order to enhance the readability of the content. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

1. Sun Y, Lin L, Tang D. Modeling mention, context and entity with neural networks for entity disambiguation. In: Twenty-fourth International Joint Conference on Artificial Intelligence. 2015.
2. Shijia E, Xiang Y. Chinese named entity recognition with character-word mixed embedding. In: Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM). 2017. p. 2055–2058.
3. Joulin A, Grave E, Bojanowski P, Mikolov T. Bag of tricks for efficient text classification. ArXiv arXiv:1607.01759. 2016.
4. Zhang S, Xu X, Pang Y, et al. Multi-layer attention based CNN for target-dependent sentiment classification. Neural Process Lett. 2020;51:2089–103.
5. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473. 2014.
6. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013.
7. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst. 2013;26.
8. Xu JWL. A comparative study on the construction of English words and Chinese characters. J Educ Inst Jilin Province. 2012;28:117–9. https://doi.org/10.16083/j.cnki.1671-1580.2012.10.008.
9. Chen X, Xu L, Liu Z, Sun M, Luan H. Joint learning of character and word embeddings. In: Twenty-fourth international joint conference on artificial intelligence. Citeseer. 2015.
10. 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. p. 981–986.
11. Yu J, Jian X, Xin H, Song Y. Joint embeddings of Chinese words, characters, and fine-grained subcharacter components. In: Palmer M, Hwa R, Riedel S, editors. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics. Copenhagen, Denmark: 2017. p. 286–291.
12. Su T-R, Lee H-Y. Learning Chinese word representations from glyphs of characters. arXiv preprint arXiv:1708.04755. 2017.
13. Zhang Y, Liu Y, Zhu J, Zheng Z, Liu X, Wang W, Chen Z, Zhai S. Learning Chinese word embeddings from stroke, structure and pinyin of characters. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019. p. 1011–1020.
14. Yang Q, Xie H, Cheng G, Wang FL, Rao Y. Pronunciation-enhanced Chinese word embedding. Cogn Comput. 2021;13:688–97.
15. Binyong YFM. Chinese romanization: pronunciation & orthography. Beijing, China: Sinolingua; 1990.
16. Li Y, Li W, Sun F, Li S. Component-enhanced Chinese character embeddings, arXiv preprint arXiv:1508.06669. 2015.
17. Xu J, Liu J, Zhang L, Li Z, Chen H. Improve Chinese word embeddings by exploiting internal structure. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies. 2016. p. 1041–1050.
18. Cao S, Lu W, Zhou J, Li X. cw2vec: learning Chinese word embeddings with stroke n-gram information. In: Proceedings of the AAAI conference on artificial intelligence. 2018. vol. 32.
19. Ma B, Qi Q, Liao J, Sun H, Wang J. Learning Chinese word embeddings from character structural information. Comput Speech Lang. 2020;60.
20. Sarzynska-Wawer J, Wawer A, Pawlak A, Szymanowska J, Stefaniak I, Jarkiewicz M, Okruszek L. Detecting formal thought disorder by deep contextualized word representations. Psychiatry Res. 2021;304: 114135.
21. Yang M, Xu J, Luo K, Zhang Y. Sentiment analysis of Chinese text based on Elmo-RNN model. In: Journal of Physics: Conference Series. Vol. 1748, IOP Publishing. 2021. p. 022033.
22. Liu B, Guan W, Yang C, Fang Z. Effective method for making Chinese word vector dynamic. J Intell Fuzzy Syst. 2023;1–12. (Preprint)
23. Huang S, Zhai Y, Long X, Jiang Y, Wang X, Zhang Y, Xie P. DAMO-NLP at NLPCC-2022 task 2: knowledge enhanced robust NER. 2022;13552:284–293.
24. Qiu Y, Li H, Qu Y, Chen Y, She Q, Liu J, Wu H, Wang H. Dureader_retrieval: a large-scale Chinese benchmark for passage retrieval from web search engine. ArXiv arXiv:2203.10232. 2022.
25. Li Z, Zhang X, Zhang Y, Long D, Xie P, Zhang M. Towards general text embeddings with multi-stage contrastive learning. arXiv:2308.03281. 2023.
26. NetEase Youdao I. Bcembedding: bilingual and crosslingual embedding for rag. 2023.
27. Wang Yuxin Hs, Sun Q. M3e: Moka massive mixed embedding model. 2023.
28. Shi X, Zhai J, Yang X, Xie Z, Liu C. 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 (Volume 2: Short Papers). 2015. p. 594–598.
29. Vaswani A. Attention is all you need. Advances in Neural Information Processing Systems. 2017.
30. Zhang Y, Liu Y, Zhu J, Wu X. FSPRM: a feature subsequence based probability representation model for Chinese word embedding. IEEE/ACM Trans Audio Speech Lang Process. 2021;29:1702–16.
31. Sukthanker R, Poria S, Cambria E, Thirunavukarasu R. Anaphora and coreference resolution: a review. Inf Fusion. 2020;59:139–62.
32. 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.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.