Six-Writings multimodal processing with pictophonetic coding to enhance Chinese language models[*]

Li WEIGANG[‡1], Mayara Chew MARINHO[1], Denise Leyi LI[2], Vitor Vasconcelos DE OLIVEIRA[1]

[1]Department of Computer Science, University of Brasilia, Brasilia 70910-900, Brazil

[2]Faculty of Economics, Administration, Accounting and Actuaries, University of Sao Paulo, Sao Paulo 05508-010, Brazil

E-mail: weigang@unb.br; mayarachewm@gmail.com; denise.leyi@gmail.com; vasconcelos.oliveira@aluno.unb.br

Abstract: While large language models (LLMs) have made significant strides in natural language processing (NLP), they continue to face challenges in adequately addressing the intricacies of the Chinese language in certain scenarios. We propose a framework called Six-Writings multimodal processing (SWMP) to enable direct integration of Chinese NLP (CNLP) with morphological and semantic elements. The first part of SWMP, known as Six-Writings pictophonetic coding (SWPC), is introduced with a suitable level of granularity for radicals and components, enabling effective representation of Chinese characters and words. We conduct several experimental scenarios, including the following: (1) We establish an experimental database consisting of images and SWPC for Chinese characters, enabling dual-mode processing and matrix generation for CNLP. (2) We characterize various generative modes of Chinese words, such as thousands of Chinese idioms, used as question-and-answer (Q&A) prompt functions, facilitating analogies by SWPC. The experiments achieve 100% accuracy in answering all questions in the Chinese morphological data set (CA8-Mor-10177). (3) A fine-tuning mechanism is proposed to refine word embedding results using SWPC, resulting in an average relative error of $\leq 25\%$ for 39.37% of the questions in the Chinese wOrd Similarity data set (COS960). The results demonstrate that SWMP/SWPC methods effectively capture the distinctive features of Chinese and offer a promising mechanism to enhance CNLP with better efficiency.

Key words: Chinese language model; Chinese natural language processing (CNLP); Generative language model; Multimodal processing; Six-Writings

1 Introduction

Chinese language and its associated culture form an essential part of human civilization's development (Wang L, 1959; Xu S, 1997; Zhao YR, 2017).

The journey from Chinese information to Chinese natural language processing (CNLP) and finally to Chinese language intelligence processing (CLIP) has been a long and historical one. Over several generations of relentless efforts, remarkable achievements have been made (Li BA et al., 2005; Feng, 2012). However, given the challenges arising from the advancements in artificial intelligence (AI), it is essential to establish a strong connection between CNLP and artificial general intelligence (AGI), as emphasized by Weigang et al. (2022). While large language models (LLMs) have demonstrated significant

success in natural language processing (NLP) (Schulman et al., 2022; Zhou J et al., 2023), they continue to face the difficulties in achieving expected results when dealing with certain intricate Chinese processing tasks. This observation is especially significant in the development of Chinese language models. Several key factors contribute to these existing gaps, making it difficult to accurately capture the nuanced expression features of the Chinese language.

1. The field of computer information science and technology, including most AI, machine learning (ML), LLMs, and image processing models, has been developed based on English and corresponding coding. Chinese language is essentially translated into a form that modern computers can understand.

2. Chinese informatization requires strict differentiation of Chinese character codes, such as the national standard GB 18030-2022 (Standardization Administration of the People's Republic of China, 2022) or Unicode (The Unicode Consortium, 2022). These codes serve as the foundation for Chinese informatization but are insufficient for effectively calculating the similarity between Chinese characters. To achieve accurate similarity results, CNLP needs specialized coding approaches and appropriate language models that can effectively comprehend Chinese characters and multimodal patterns.

3. The lack of a standardized Chinese character coding suitable for Chinese language models has resulted in the absence of (1) a unified presentation for prompting questions and (2) a standardized method for calculating the similarity between Chinese characters. Despite most research findings indicating high similarity, CNLP's actual results are less than ideal.

In this context, an important challenge—and the main goal of this study—is to design an appropriate Chinese multimodal processing and coding system, which allows ML models to effectively master Chinese characters and multimodal patterns, while improving the accuracy of various CNLP tasks.

When reflecting on the history of development of the Chinese language, the contributions of Shen XU, a philologist during the Eastern Han Dynasty, are worth noting. He compiled the earliest Chinese dictionary in history called "*Shuo Wen Jie Zi*" (说文解字) (Xu S, 1997). This reference book systematically analyzed the visual forms of Chinese characters and provided detailed explanations of their origins (Song et al., 2006). The methods introduced for character creation and usage are collectively referred to as "Six-Writings" (六书). The classification system in this cited book categorizes Chinese characters into six types according to their multimodal composition and usage: pictograms, phono-semantic characters, simple ideograms, compound ideograms, transfer, and loan characters (Yeromiyan, 2022).

This paper introduces a novel approach called Chinese Six-Writings multimodal processing (SWMP) for Chinese language models, as shown in Fig. 1. It is referred to as the Six-Writings concept for phono-semantics, pinyin, image, audio/video, property, and understanding. Within this unified coding framework, character and word coding can be seamlessly integrated with Chinese grammar while retaining flexible Chinese customary characteristics. The paper presents the proposal and applications on Six-Writings pictophonetic coding (SWPC), the first part of SWMP. It puts forward the basic concepts and methods for text/image processing based on character generation and representation of prompt in addition to outlining future research directions.

With the development of SWPC, let us examine the question-and-answer (Q&A) mode concerning standard Chinese word formats. In this context, we can consider a word pair (护佑bless, 佑护bless) and predict their similarity. Manual evaluation using the reference materials indicates a similarity score of 0.883 (Huang JJ et al., 2019), while calculation using the language model yields a score of 0.634 (Jin et al., 2022). To calculate this, we apply our SWPC approach, taking into account the Chinese word formation method's out-of-order word rules (AB = BA). The similarity of this word pair is determined as follows: $\mathrm{Sim}(f(\mathrm{AB}), f(\mathrm{BA})) = \max(\mathrm{Sim}(f(\mathrm{AB}), f(\mathrm{BA})), \mathrm{Sim}(f(\mathrm{AB}), f(\mathrm{AB}))) = 1.000$, $f(\cdot) = \mathrm{SWPC}(\cdot)$. Note that >90% of modern Chinese characters possess pictophonetic features (Zhang B, 2008). We present some calculation modes for the pictophonetic coding of other Chinese phrases, such as AQAB, AABB, and ABAB (Zhang B, 2008; Jin et al., 2022), which effectively reflect the nuances of CNLP. These modes bear similarities to the Q&A style used in prompt engineering (Liu PF et al., 2023), which tackles the challenges of "pretrain, prompt, and predict," commonly encountered with language models.

Pictophonetic		Pinyin		Property	Image	Audio/Video	Understanding
イ 57	白 5311	bó băi	2 3	Gender(性别), age(年龄), etc.	伯		Word embedding (词嵌入向量)
形声		拼音/声调		说文	图像	音频/视频	会意

Fig. 1 Framework of Six-Writings multimodal processing (SWMP) for Chinese characters

To test our proposed methods, we design some analogical reasoning models (in which SWPC adheres to the grouping rules of Chinese words) and study some experimental situations, including processing of the Chinese analogical (CA8) data set (Li S et al., 2018). The analogical reasoning method answers all questions of the CA8-morphological (CA8-Mor-10177) data set and 12.05% of the questions in the CA8-semmantic (CA8-Sem-7636) data set with 100% accuracy. The SWPC approach can also be used to fine-tune the results of word embedding and analyzes 39.37% question pairs in the Chinese wOrd Similarity-960 (COS960) data set (Huang JJ et al., 2019). The relative error between similarity calculation and artificial basis evaluation is $\leq 25\%$. The analysis shows that our proposal embodies the delicate local representation and overall relevance of the Chinese language, and has the potential to supplement current CNLP theory and technology. Our study makes significant contributions in the following aspects:

1. The variation problem in Chinese similarity calculation is revealed. Further, we propose an augmentation method for converting Chinese characters (and pinyin) from letter code to numerical code for CNLP, as shown in Section 3.

2. We construct a framework called SWMP, which allows for the unified multimodal coding of Chinese characters (and words) with appropriate granularity for language models in Section 4.

3. We introduce a novel method SWPC and its applications, establishing a generative mechanism for representing characters (and words) using pictophonetic features, as shown in Section 5.

4. An experimental database is established, consisting of images and SWPCs for the root, radical, and components, so that characters can be processed in dual mode through SWPC and graphics, as shown in Section 6.

5. We have developed the analogical reasoning modes and prompting functions of SWPC for various word combinations, such as repetition (AABB), prefix (PQAB), and suffix (ABPQ). Further, SWPC is successfully applied to the CA8 data set (Li S et al., 2018) and COS960 data set (Huang JJ et al., 2019), as shown in Sections 7 and 8.

6. We point out the research directions for the future, as shown in Section 9.

2 Related works

To improve the performance of CNLP, several methods have been introduced to effectively capture the semantic and morphological information of Chinese characters. Chen XX et al. (2015) proposed a framework to create low-dimensional distributed word embeddings, named the character-enhanced word embedding (CWE) model, which includes subword information and character n-grams. Another framework for the joint learning of character and word embeddings uses a convolutional neural network (CNN) to model the internal structure of each character and a recurrent neural network to model the context of each word (Xu J et al., 2016). Yu et al. (2017) proposed an approach named the joint learning word embedding (JWE) model to jointly embed Chinese words, their characters, and fine-grained subcharacter components.

Cao et al. (2017, 2018) proposed cw2vec, which decomposes Chinese characters or words into strokes, reorganizes them with semantic and morphological levels through n-gram, and learns word embedding. Additionally, character embedding models, including pinyin and Wubi (WB) input letter coding (The Wubi Group, 2000), were developed to capture structural and phonetic information (Zhou JN et al., 2019). Zhuang et al. (2019) constructed the vector representation of strokes and adjacent strokes as sub-character embedding, combined with character embedding to achieve continuous enhancement of word embedding. Cj2vec (Kang et al., 2019) uses the Cangjie alphabet code to represent Chinese

characters; it trains the model in the same way as cw2vec. Zhang Y et al. (2019) introduced ssp2vec, which integrates strokes, structures, and pinyin of Chinese characters into a superset to capture the characteristics of Chinese sub-characters. Wang SR et al. (2020) introduced radical and stroke-enhanced word embeddings (RSWE), which captures the local structural features of words and integrates implicit information from radicals to enhance semantic embedding. Zhao DP et al. (2021) conducted a study on the calculation of Chinese text similarity using the sequence alignment algorithm. Learning from Chinese word embeddings, Lu et al. (2022) used the semantic relevance hidden in lower components (sub-characters) to further distinguish the semantics of corresponding higher components (characters and words). Jin et al. (2022) developed a model that captures semantic and morphological information using the four-corner (FC) features of characters and words. More recently, based on graph clustering, sense embedding was used to improve entity linking in the Chinese domain (Zhang ZB et al., 2023).

Some valuable research works have used images of the Chinese characters to improve CNLP tasks. The character glyph features are directly learned from the bitmaps of characters by convolutional auto-encoder (convAE), and the glyph features improve Chinese word representations (Su and Lee, 2017). Glyce was developed to enrich the pictographic evidence in characters and increase the model's ability to generalize (Meng et al., 2019). A multimodal model, Glyph2Vec, was developed to extract visual features from word glyphs to expand current word embedding space for out-of-vocabulary (OOV) word embedding (Chen HY et al., 2020). Related to the study of Chinese character glyphs, some recent advances in image synthesis and generation using soft shadow networks are particularly interesting (Sheng et al., 2021, 2022, 2023). Specifically, the approach of "pixel height," which is a new geometric representation that encodes the correlations among objects, ground, and camera posture, is a novel idea and is of significant reference value for the further synthesis and generation of Chinese characters using text/image component coding.

Notably, Song et al. (2006) studied the productive expression of the phonetic semantic relations of *Shuo Wen Jie Zi* (说文解字). Wang JT (2011) explored the calculation of Chinese string similarity based on the clustering features of Chinese characters. Liu MD and Liang (2021) proposed a method of dividing Chinese characters into squares based on the calculation of similarity using radical knowledge representation learning and thereby constructed the radical knowledge graph spectrum. These works provide valuable insights and raise the issue of calculation of similarity between Chinese characters. Actually, most CNLP tasks involve calculating similarity, but there is still considerable room for improvement in specific calculations, warranting further discussion.

The aforementioned research essentially integrates the features of Chinese characters into word embedding technology to improve CNLP. However, some studies have pointed out that these attempts have limited effectiveness and may introduce unnecessary noise due to nonstandard representation of Chinese characters (Jin et al., 2022). LLMs have different processing costs according to the different languages processed, with English having the lowest processing cost (Petrov et al., 2023). Another widely recognized problem in word embedding technology is the vector dimension of high-frequency words, which usually ranges from 50 to 300. In other words, the dimension is inefficient and costly. Lengthy digital codes for high-frequency characters are not efficient at all. Taking Chinese as an example, there are >1000 commonly used characters, so it is very important to use effective and accurate character encoding to reflect the unique features of the Chinese language.

Our research distinguishes itself from previous studies by incorporating a multimodal analysis of Chinese characters. While many previous studies have relied solely on the morphological and pictophonetic features for word embedding, our research considers the Six-Writings of Chinese characters and uses a generative radical/component coding approach. This mechanism provides a more comprehensive understanding of Chinese character representation, which is essential for accurately calculating the similarity between Chinese characters. This paper presents a thorough investigation of the challenges associated with this task and proposes a new approach that enhances the capabilities of CNLP.

3 Variation in similarity calculation and augmentation methods

This section introduces the concept of statistical coefficient of variation (Cv). It further explores the variation issues and solutions related to numerical codes such as WB and pinyin.

3.1 Concept of the coefficient of variation

In the probability theory and statistics, when comparing the similarity (dispersion) between two sets of data, variations can occur if the measurement scales or data dimensions of the two sets are inconsistent. Cv is a normalized measure of the degree of dispersion in a probability distribution. It is defined as the ratio of the standard deviation to the mean (Everitt and Skrondal, 2010):

$$\mathrm{Cv} = \frac{\sigma}{\mu}, \tag{1}$$

where Cv is the coefficient of variation, μ is the mean value, and σ is the standard deviation.

In the field of NLP, numerous tasks require the computation of similarity between characters or words. In the case of English and other alphabetic languages, American Standard Code for Information Interchange (ASCII) codes, which align with computer information processing, are commonly used for digitizing letters. Subsequently, the similarity between words or sentences can be determined. It is also suggested that cosine similarity is well-suited for essential language modeling tasks in the literature (Turney, 2012; Mikolov et al., 2013).

Currently, in CNLP research, there are two primary technical approaches aimed at improving the understanding of Chinese by language models through the utilization of Chinese character features: (1) integrating Chinese character representations, such as letter or numerical codes, during the word embedding process to acquire the necessary character or word representation vectors for the language model; (2) using Chinese characters or numerical codes directly to construct Chinese knowledge maps and subsequently determining character similarity.

These two techniques typically use pinyin, strokes, FC numbers, WB letter, or numerical codes to represent Chinese characters and words. The similarity between characters or words is then calculated to accomplish various CNLP tasks. However, when training ML models to embed vectors for words, there is no standardized input mode or coding for the digital representation of Chinese characters. The stroke of Chinese characters is directly digitized (usually between 1 and 5) and inputted into the language model (Cao et al., 2017, 2018; Wang SR et al., 2020). Others are represented using letter codes through Chinese character input methods such as Cangjie (Kang et al., 2019), WB (Zhou JN et al., 2019), or FC numbers (Jin et al., 2022).

Due to the diverse techniques used to describe the features of Chinese characters and the varying granularities of the language elements in these representation methods, variations in similarity calculations are inevitable. These are the challenges that CNLP is facing. We use the concept of the statistical Cv to describe the variation problem in calculating the similarity between characters, and propose the solutions of normalization and augmentation to deal with these variations.

3.2 Augmentation of WB code

WB code is a widely used Chinese input method for entering characters on computers or mobile devices (The Wubi Group, 2000). The term Wubi translates to "five-stroke typing," referring to the fact that each Chinese character is constructed using five basic strokes. The WB input method is based on the principle of inputting characters according to their shape, rather than their pronunciation. Users can enter the character by typing its shape, resulting in faster and more accurate input compared to pronunciation-based methods.

On a traditional QWERTY keyboard, the WB input method organizes the arrangement of multiple-digit roots to correspond to specific keys, following the composition rules of Chinese characters (The Wubi Group, 2000). It is arranged in the following manner: the keyboard is divided into five zones, numbered 1–5, and each zone has five keys. These keys are assigned respective letter keys, also numbered 1–5, arranged from the middle to the outer ends of the keyboard. By combining the zone and position numbers, 25 codes are formed. The zone codes are as follows: 11(G), 12(F), ..., 51(N), 52(B), ..., 55(X). Table 1 provides a comparison between the letter codes (key positions) and the number codes (partitions) of WB on the QWERTY keyboard, in addition to a statistical analysis of their variability.

According to the rules of WB, although WB

letter codes are used to encode Chinese characters, a corresponding WB number code is naturally formed for each character. These codes range from 11 to 55, with an average value (μ) of 33, a standard deviation (σ) of 14.5057, and a Cv of 0.4396. Because of the narrow value range of numerical codes, the result of cosine similarity calculations is obviously larger. For example, when calculating the similarity of the word pair (粘膜mucous, 黏膜membrane) in the COS960 data set (Huang JJ et al., 2019), the benchmark similarity in the original data set is 0.6500 (2.6/4.0). However, the cosine similarity calculated using the WB number code is 0.8867, resulting in a relative error of 36.42% compared to the benchmark.

To address this variation, data augmentation is applied by expanding the value range of WB number codes from 11–55 to 11–99 according to the rules of WB. The average corrected number for the 25 codes becomes 55, with σ being 29.0115 and Cv being 0.5275 (Table 1). The advantage of the augmentation is that it enlarges the distance between characters in numerical coding, thus producing relatively small cosine similarity values. For example, the cosine similarity of the aforementioned word pair (粘膜mucous, 黏膜membrane) becomes 0.7997, with a relative error of 23.03% compared to the benchmark similarity.

Table 1   Coefficient of variation (Cv) for WB letter/numerical codes and augmented codes

WB Letter/Number		AWB Letter/Number	
G/11	W/34	G/11	W/57
F/12	Q/35	F/13	Q/59
D/13	Y/41	D/15	Y/71
S/14	U/42	S/17	U/73
A/15	I/43	A/19	I/75
H/21	O/44	H/31	O/77
J/22	P/45	J/33	P/79
K/23	N/51	K/35	N/91
L/24	B/52	L/37	B/93
M/25	V/53	M/39	V/95
T/31	C/54	T/51	C/97
R/32	X/55	R/53	X/99
E/33		E/55	
μ	33	μ	55
σ	14.506	σ	29.011
Cv	0.4396	Cv	0.5275

AWB: augmented WB

Based on the above analysis, it is evident that the numerical representation of different scales leads to different calculation results when using similarity measures such as cosine similarity. This variation in similarity calculation highlights the problem of variation in Chinese character similarity, which warrants attention. Notably, if Hamming (HM) similarity is used, the data augmentation in Table 1 does not affect the similarity calculation. The HM similarity of WB and the normalized and augmented coding (AWB) for the aforementioned word pair (粘膜mucous, 黏膜membrane) both yield a similarity score of 0.5625, with a relative error of 13.46% compared to the benchmark.

The FC method has been used to index and search Chinese characters based on their shapes. Each Chinese character has four corners numbered from zero to nine, representing the basic elements of the character's shape, such as horizontal strokes, vertical strokes, or dots. By combining the numbers of these basic elements, a four-digit (plus one complement number) code is created to identify the characters. Recently, FC numbers have been used to represent Chinese characters, aiming to improve the effectiveness of CNLP (Jin et al., 2022).

We conduct statistical analysis on the FC number table, which comprises 27 585 Chinese characters (https://github.com/zongxinbo/rime-zong/blob/master/sjhm.dict.yaml), yielding the following insights: 18.30% (5047) of FC numbers are reused, and 79.42% (21 909) Chinese characters share the same FC number with other characters. In certain cases, 117 Chinese characters (such as 带 (bring) and 芳 (fragrant)) simultaneously apply the FC number 44227. The Cv for the whole set of 27 585 characters is 0.5286.

Furthermore, we examine the WB letter coding table, consisting of 27 536 Chinese characters (https://github.com/yanhuacuo/98wubi-tables/blob/master/GB18030-27533.txt), and obtained the following findings: 15.34% (4224) of WB letter codes are reused, and 35.05% (9650) Chinese characters share the same WB letter code with other characters. There are 21 instances of Chinese characters (e.g., 殼 (shell)) that use the WB code FPGC concurrently. The Cv for the whole set of 27 536 characters is 0.6185.

Table 1 lists the specific normalized and augmented WB (AWB) codes that serve two purposes. First, we propose a coding approach to combine FC numbers and WB in order to effectively reduce the occurrence of duplicate codes. Second, the number

range of the FC number is 0–9, and the number coding of WB is augmented from 11–55 to 11–99, which does not affect the coding rules of WB. Table 1 and the above discussions also show how normalized and augmented coding can improve the Cv, resulting in consistent similarity calculation between characters (or words) compared to using just WB or FC numbers.

3.3 Augmentation of pinyin numerical code

Chinese pinyin consists of initials and vowels, with a total of 23 initials—b p m f d t n l g k h j q x zh ch sh r z c s y w, and 24 vowels—a o e i u ü ai ei ui ao ou iu ie üe er an en in un ün ang eng ing ong. By using the lowercase letter v instead of ü, all the 26 English letters can fully represent the initials and vowels. Referring to the English alphabet coding of the GB 18030-2022 standard, the numerical codes of the capital letters A–Z range from 65 to 90 (decimal). The average value (μ) is 77.50, σ is 7.3598, and Cv is 0.0950, as shown in Table 2 (because of the regularity of the sequence, several rows are omitted). This small Cv affects the calculation of phonetic similarity between Chinese characters. Similarly, if only lowercase alphanumeric codes are used to represent the initials and vowels (a–z: 97–122), a similar situation arises: μ is 109.50, σ is 7.3598, and Cv is 0.0672.

To address the problem of the small Cv, we propose a normalization and augmentation method of digital conversion for pinyin. Specifically, the initial letters are coded using uppercase letters from GB 18030-2022, the vowel letters are coded using lowercase letters, and the vowel letter ü is coded using the lowercase letter v. Table 2 presents the phonetic alphabet coding and a comparison between the GB and augmented codes.

The first column represents the 26 uppercase and lowercase English letters. The second column shows the decimal codes in GB 18030-2022 corresponding to these uppercase and lowercase letters. For example, the code for A/a is 65/97. Column 3 normalizes the codes of the 52 letters A–Z and a–z from 65–122 to 11–99, using integers. In this process, the code for F is adjusted from 70 to 19, f is adjusted from 102 to 68, and n is adjusted from 110 to 81 to avoid using 0 as a code. After normalization and augmentation, the coding for A/a becomes 11/59. The average value of the 52 letter codes is 54.8669, σ is 27.4341, and Cv is 0.5000. It is close

Table 2　Coefficient of variation (Cv) for the GB code and augmented codes to present letters

Letter	GB code	Augmented code
A/a	65/97	11/59
B/b	66/98	13/62
C/c	67/99	14/63
D/d	68/100	16/65
E/e	69/101	17/66
F/f	70/102	19/68
...	...	...
V/v	86/118	43/93
W/w	87/119	45/94
X/x	88/120	46/96
Y/y	89/121	48/97
Z/z	90/122	49/99
Cv of GB A–Z	26 letters	0.0950
Cv of GB a–z	26 letters	0.0672
Cv of augmented A–Z and a–z	52 letters	0.5000

to 0.5275, the coefficient of the augmented number code in Table 1.

Based on the normalized and augmented Chinese pinyin numerical coding presented in Table 2, the coding for the pinyin zh ā ng of 张 (zhang) is as follows: the code for the initial consonant ZH is 4922, and the code for the vowel ang is 598169. To analyze the pronunciation similarity between 张 (zhang) and 黄 (huang), we calculate the cosine similarity using the GB numbers of capital letters. The similarity of cosine by pronunciation between the two characters is 0.9752. However, when the augmented GB numbers of uppercase and lowercase letters are used in Table 2, the similarity of pronunciation cosine between the two characters is 0.9251. In both cases, the HM similarity is 0.7.

This example emphasizes once again that there are differences in calculating the cosine similarity for Chinese characters due to the different numerical codes assigned to letters. The proposed pinyin numerical augmentation method is very similar to the conversion between English letters and numbers. In Chinese pinyin, we use uppercase letters/numbers to represent initials and lowercase letters/numbers to represent vowels, thus increasing the numerical distance between initials and vowels. This method has a positive effect on calculating the cosine similarity between pinyin representations.

4 Framework of SWMP

This section provides an overview of the Six-Writings concept and presents the framework of SWMP for Chinese characters (or words), and the related discussion.

4.1 Concept of Six-Writings

Modern Chinese characters have evolved from ancient Chinese characters, and understanding the structure of Chinese characters requires knowledge of their ancient counterparts. Shen XU's *Shuo Wen Jie Zi* (说文解字) from the Eastern Han Dynasty classified Chinese characters into Six-Writings (六书, liùshū), representing six different structural types (Xu S, 1997; Yeromiyan, 2022):

1. Pictograms (象形): characters that visually resemble the objects they represent, such as 日 (Sun).

2. Ideograms (指事): characters wherein the form itself expresses a specific meaning, representing abstract concepts such as numbers or directions, e.g., 上 (up) and 下 (down).

3. Compound ideograms (会意): characters composed of two or more ideograms, conveying complex meanings, e.g., 亻 (person)+ 木 (wood)=休 (rest).

4. Pictophonetic (形声): characters composed of phonetic components for pronunciation and semantic components for meaning, such as 羊 (sheep), 洋 (foreign), and 样 (kind), all pronounced as "yang."

5. Transformed cognates (转注): characters that have undergone changes in pronunciation over time and no longer resemble their original form, for example, 考 (old)=老 (old).

6. Loan characters (假借): words borrowed from other languages that share similar pronunciation or meaning, e.g., 乎 (hu)=呼 (call).

The Six-Writings of Chinese characters is a multimodal way of understanding Chinese characters, which was summarized by the ancients. It takes into account how human beings understand Chinese characters through their five senses. Following *Shuo Wen Jie Zi*, this paper proposes a comprehensive multimodal processing framework and coding method for Chinese characters (or words), which adapt to modern Chinese language models.

4.2 Framework of SWMP

The SWMP framework for Chinese language models consists of six parts: (1) pictophonetic, (2) pinyin, (3) property, (4) image, (5) audio/video, and (6) understanding (word embedding), as shown in Fig. 1.

The first part deals with the letter/number coding of the pictophonetic features of the radical and components of Chinese characters. The second part is pinyin and tone coding of Chinese characters. The third part includes coding the main attributes of characters, such as human-related attributes, gender, and semantic interpretation. The fourth part represents image information related to characters. The fifth part contains the audio/video information. Finally, the sixth part deals with the word embedding vectors and other semantic information of characters (or words) for understanding. For instance, the character 伯 (uncle) is used as an example in the image, and its WB letter code is WRG, with the corresponding number code being 343211. According to Table 1, the AWB code is 575311.

This comprehensive multimodal processing framework allows for a detailed representation of Chinese characters and words in the Six-Writings style, facilitating various language processing tasks and enabling a deeper understanding of their structure and attributes.

In practical applications, different parts of the SWMP framework can be combined based on specific task requirements. For example, Section 5 of this paper focuses on the detailed explanation of the first part of SWMP for pictophonetic representation. The second part creates the Six-Writings pinyin code, while the first and third parts can be combined to form the Six-Writings semantic code. Furthermore, integrating the first, second, and fourth parts results in the comprehensive text/pronunciation/image processing of Six-Writings.

To facilitate the analysis, this paper provides the basic naming specifications for Chinese coding according to *Xinhua Dictionary*, including the following:

1. The FC numbers consist of a total of five digits, each ranging from zero to nine.

2. Stroke code (SC) varies in length depending on the complexity of Chinese characters, ranging from one to five.

3. The WB numerical code comprises a total of eight digits. If the code has <8 digits, zeros are appended to the end.

4. The combination of AWB and FC code is known as SWPC, which usually uses a total of 10 digits.

5. Six-Writings pinyin (SWPY) is a numerical code that represents pinyin, including 11 digits for the pinyin and tone codes.

6. Six-Writings image code (SWIC) is a digital representation of a character by an image, typically represented by a 0–1 digital matrix of Chinese characters.

In the following we provide an introduction to the rules, methods, and applications associated with the combination and utilization of SWMP. We primarily focus on the overall design framework of SWMP; it places particular emphasis on the research methods and applications of SWPC, SWPY, and SWIC. The coding and application of audio/video and property understanding by multi-dimensional classification (Saleh and Weigang, 2023) will be addressed as future research endeavors.

4.3 SWPY code

The digital conversion and augmentation of Chinese pinyin are shown in Table 2. The pinyin tone represents the tone in Putonghua, commonly known as the four tones. The level tone (first tone) is represented by ¯, such as lā; the rising tone (second tone) is represented by ´, such as lá; the falling-rising tone (third tone) is represented by ˇ, such as lǎ; the falling tone (fourth tone) is represented by ˋ, such as là. Since the longest pinyin does not exceed five letters, such as zhang (zhāng), each letter is converted into two numbers along with the tone. Therefore, the numerical code of pinyin consists of 11 digits. For example, the pinyin (zhāng) for character 张 (zhang) is coded as 49225981691, and the pinyin (lǜ) for character 绿 (green) is coded as 28930000004.

In summary, modern Chinese characters are characterized by flexibility and diversity due to historical evolution and other factors. This includes the presence of polysemy (multiple meanings) and homophony (multiple pronunciations). Recognizing Chinese characters requires the synthesis of multimodal information like Six-Writings, such as pictophonetic codes, pinyin, images, and others. The proposed SWMP for CNLP is a promising approach for achieving this goal.

5 SWPC approach

This section introduces the SWPC of Chinese characters and its application in measuring the similarity between the characters or words.

5.1 Pictophonetic property

Chinese characters are logograms with a hierarchical structure consisting of strokes and components. Strokes are the basic units of character formation, and components are formed by combining strokes (Wang L, 1959; Wang SK, 2016). It is estimated that over 90% of Chinese characters belong to the phono-semantic (形声, xíngshēng) category (Zhang B, 2008; Yeromiyan, 2022). Phono-semantic characters are typically composed of a semantic component, namely the radical (形旁, xíngpáng) and a phonetic component (声旁, shēngpáng). The semantic component represents the overall meaning or category of the character and is commonly referred to as the radical component. The phonetic component, on the other hand, serves as a means to differentiate characters with similar meanings or pronunciations. According to Fig. 1, the first part of SWMP is the pictophonetic code for Chinese characters.

In this study, we propose a new pictophonetic code, SWPC, which is derived from the AWB numeric codes and FC numbers. Chinese pictophonetic features are considered as combinations of radical and phonetic components, which can be further categorized into the following eight types (Wang SK, 2016): (1) radical on the left, phonetic on the right; (2) phonetic on the left, radical on the right; (3) radical on the top, phonetic on the bottom; (4) radical on the bottom, phonetic on the top; (5) radical on the inside, phonetic on the outside; (6) radical on the outside, phonetic on the inside; (7) radical occupying a corner; (8) phonetic taking a corner. Based on these eight classifications, SWPC consists of the radical code (semantic component) and the phonetic code (phonetic component), as follows:

1. SWPC radical code consists of AWB radical code (two digits) and FC radical code (two digits). The phonetic code is composed of the WB component code (six digits). Generally, it is composed of 10 digits.

2. If a character has a four-digit WB shaped side code, the radical code of SWPC is the same as four-digit code. The phonetic code is formed by combining the remaining WB component codes (four-digit numbers) with FC component codes (two-digit numbers), resulting in a total of 10 digits.

3. For other Chinese characters that do not possess pictophonetic properties, their codes are based on the AWB numeric code (eight digits, with trailing zeros if necessary) followed by the last two digits of the FC numbers, totaling 10 digits.

SWPC is expected to be a more comprehensive and informative representation of Chinese characters than previous pictophonetic codes. It has the potential to improve the accuracy of Chinese character recognition and other NLP tasks. The following subsections provide detailed explanations and examples of various combinations to form SWPC. For specific principles of stroke splitting and coding rules, please consult the relevant resources on the WB method (The Wubi Group, 2000).

5.2 Character by left and right components

Most Chinese characters are formed by combining the left and right components. For example, character 横 (horizontal) has the WB letter code SAMW and the WB numeric code 14152534. According to the WB coding rule, as shown in Fig. 2, the key position of the left radical 木 (wood) on the QWERTY-based keyboard is S, and its position code is 14. The code of the phonetic component 黄 (yellow) is 152534. According to Table 1, the normalized and augmented WB number is 17193957, with a radical code of 17 and a phonetic code of 193957, totaling 8 digits.

Fig. 2 WB coding of character 横 (horizontal) (http://life.chacuo.net/convertxuezi)

Another character 酮 (ketones) has the WB code SGMK, the WB numeric code 14112523, the normalized and augmented code 17113935, the radical code 17, and the component code 113935. The radical codes of 横 (horizontal) and 酮 (ketones) are the same, i.e., 17, resulting in duplicate codes. Note that the FC numbers of these two characters are different. The FC number of character 横 (horizontal)

is 44986, the radical 木 (wood) has a number code of 49, and the phonetic 黄 (yellow) has a number code of 486. To avoid code duplication and enhance the digital representation of the pictophonetic features, we combine the advantages of WB and FC and modify the radical 木 (wood) of character 横 (horizontal) to 1749, while keeping the phonetic 黄 (yellow) unchanged at 193957. The combination of the normalized and augmented WB numerical code and FC numbers is called SWPC. In this case, character 横 (horizontal) is coded as 1749193957, with a total of 10 digits.

Similarly, the FC number of character 酮 (ketones) is 17620, with the radical 酉 (unitary) having a code of 16 and the phonetic component 同 (same) having a code of 720. The SWPC is 1716113935, effectively distinguishing the two radicals 木 (wood) and 酉 (unitary). This demonstrates the advantages of using SWPC to represent Chinese characters, especially when calculating the similarity between Chinese characters (or words) and conducting multimodal text/image processing.

5.3 Character by upper and lower components

Let us take character 写 (write) as an example to explain SWPC in the case of upper and lower combination. Its WB code is PGNG, and its WB numerical code is 45115111. The radical 冖 (top) is positioned at the top. According to the WB rule (Fig. 3), the key position is P, the radical code is 45, and the augmented code is 79. The phonetic code is 115111. Referring to its FC numbers as 37127, the radical number is 37, and the phonetic number is 127. By the combination of WB and FC, the SWPC for 写 (write) is 7937119111.

Fig. 3 WB coding of character 写 (write) (http://life.chacuo.net/convertxuezi)

Since there are multiple characters with the same radical key code, such as character 宝 (precious) also related to the P key, with WB letter code PGYU and WB numeric code 45114142, considering that the FC number of character 宝 (precious) is 30103, we also combine WB with the FC numbers. The SWPC for 宝 (precious) is 7930117173,

where the radical code is 7930 and the phonetic code is 117173, avoiding duplication with the WB radical code of 写 (write).

5.4 Character by internal and external components

Let us take character 医 (medicine) as an example to explain SWPC for the internal and external combination. Its WB code is ATDI and its WB numeric code is 15311343. The radical 匚 is positioned on the outside, with WB key position A and number code 15 (Fig. 4). The phonetic component 矢 (arrow) is positioned on the inside, with code 311343. The FC number of word 医 (medicine) is 71718. The SWPC is 1977511575, with the radical code being 1977 to avoid code duplication, and the phonetic code is 511575.

Fig. 4 WB coding of character 医 (medicine) (http://life.chacuo.net/convertxuezi)

In addition to the aforementioned composition modes, there are other combinations. Specifically, we can refer to the WB rules and the methods introduced here to define their SWPC.

We have produced a database of 1000 commonly used Chinese characters using SWPC. The repetition rate of SWPC is <0.2%, which is much lower than the repetition rates of other encoding systems, such as FC numbers (24.5%) and WB code (1.5%).

5.5 SWPC for similarity between words

A morpheme is the smallest unit in a language that combines sound and meaning. For Chinese, it serves as the basic building block for word formation and is the smallest independent language unit within a sentence (Huang BR and Li, 2012).

There are two types of morphemes: roots and affixes. Roots are meaningful morphemes that can occur in different positions within a word, while affixes are adhesive morphemes with fixed positions and unreal meanings in compound words. The basic patterns of Chinese word formation include root, root+root, prefix+root, and root+suffix. The root, prefix, or suffix can consist of one or more Chinese characters. For instance, 国家 (country) and 学生 (student) are compound words formed by combining two roots, while 老子 (Laozi) is a compound word consisting of a prefix and a root, and 橘子 (orange) is a compound word consisting of a root and a suffix (Huang BR and Li, 2012).

The coding of Chinese words can be achieved by naturally combining the SWPC of characters. This is one of the generative properties of SWMP. For instance, let us consider word 朦胧 (hazy), where the WB code for character 朦 (deceive) is 33154533 (EAPE), and the FC number is 74232. Similarly, the WB numeric code for character 胧 (hazy) is 33135551 (EDXN), and the FC number is 73214. The radicals for both characters, 月 (Moon), share the SWPC radical code 3372. Therefore, the WB code for word 朦胧 (hazy) becomes 33154533 33135551, and the SWPC is 5572197955 5572159991.

Using SWPC, we can naturally deduce the calculation of similarity between words, which is important for various tasks of CNLP. Numerous studies have explored the calculation of similarity between Chinese words (Zhao DP et al., 2021). In particular, Wang JT (2011) refined the result of string matching by clustering patterns of Chinese characters, and proposed a two-level similarity model.

Table 3 gives the results of similarity calculation between the selected word pairs in Wang JT (2011). Although these results may not be the standard answers, we use the results of EE+WB and JE+WB as benchmarks. Among them, EE stands for the similarity algorithm based on the editing distance, and JE stands for the similarity algorithm based on the Jaro-Winkler distance, with the WB code being used. The WB numerical code and SWPC are used in our research.

In Table 3, the Chinese word pairs are translated as the following: (松驰slack, 松弛slack), (走漏leaked, 走露leaked), (李鹏Li Peng, 李朋Li Peng), (曝光exposure, 暴光exposure), (霎时instant, 刹时instant), (修茸repair, 修葺repair), (赋予confer, 服役service), and (虚心humility, 步履walk).

In Table 3, using FC, five-stroke, and SWPC, the similarity between cosine and HM similarity is calculated for comparison. Lines 11 and 12 in the table display the mean relative error (MRE) and variance of relative error (VRE) of these similarities compared to the results obtained with EE+WB (Wang JT, 2011). Lines 13 and 14 show the MRE and VRE of these similarities compared to those calculated using JE+WB. From the perspective of

the phono-semantic features of Chinese characters, cosine similarity is not suitable for describing the similarity between Chinese characters (and words). In all cases, the average and VREs are very high compared with the benchmark results. Note that it is impossible to distinguish word pairs with different glyphs. For instance, the cosine similarity between 虚心 (humility) and 步履 (walk) is above 0.7041. However, the calculation of HM similarity is relatively stable, especially when the SWPC (HM* column) proposed in this paper is adopted. Compared to the benchmark results of JE+WB, the MRE with JE is 0.1180, and the VRE with JE is 0.0257. Moreover, the HM similarity of the word pair 虚心 (humility) and 步履 (walk) is 0.4313, enabling a better distinction between them.

Table 3 Similarity comparison among the different code methods

Word		FC		WB		SWPC		EE+WB	JE+WB
		COS	HM	COS	HM	COS	HM*		
松驰	松弛	0.922	0.800	0.889	0.625	0.873	0.713	0.750	0.861
走漏	走露	0.774	0.600	0.825	0.500	0.742	0.550	0.500	0.500
李鹏	李朋	0.890	0.900	0.908	0.750	0.879	0.825	0.875	0.967
曝光	暴光	0.678	0.700	0.956	0.750	0.941	0.725	0.750	0.925
雾时	刹时	0.807	0.600	0.873	0.625	0.815	0.613	0.500	0.500
修葺	修茸	1.00	1.00	0.877	0.688	0.842	0.844	0.875	0.963
赋予	服役	0.719	0.400	0.756	0.125	0.675	0.263	0.125	0.250
虚心	步履	0.704	0.300	0.917	0.563	0.884	0.431	0.292	0.528
MRE with EE		0.976	0.367	1.140	0.192	1.000	0.241	–	–
VRE with EE		0.838	0.027	1.390	0.555	1.320	0.298	–	–
MRE with JE		0.478	0.232	0.546	0.189	0.467	0.118	–	–
VRE with JE		0.187	0.255	0.499	0.112	0.467	0.026	–	–

MRE: mean relative error; VRE: variance of relative error; EE: similarity algorithm based on the editing distance; JE: similarity algorithm based on the Jaro-Winkler distance

6 SWPC for text/image processing

SWPC provides convenience for multimodal processing of Chinese characters using image and text data. Drawing from the coding process of WB (The Wubi Group, 2000; Gao, 2003), tens of thousands of Chinese characters can be categorized into three main situations based on their character formation rules:

1. The character itself serves as a root character.

2. The character's root is a radical (semantic) of another Chinese character, represented by a radical code (a part of SWPC).

3. The character's root is a component (phonetic) of another Chinese character, represented by a phonetic code (a part of SWPC).

Fig. 5 illustrates some Chinese characters in the form of a matrix. The Chinese character pictures in the first and third columns are the same as those in the third row. These characters have three identities. Let us take 口 (mouth) as an example. It is a Chinese character itself (root) with an SWPC of 35353535 and an FC number of 60000. Furthermore, 口 (mouth) functions as a radical with an SWPC code of 3560, and also serves as a component with a phonetic code of 351100 by SWPC. The character 斤 (catty) shares similar features. As shown in Fig. 5, the Chinese character 听 (listen) (3560533100) is generated by combining the radical code of 口 (mouth) with the phonetic code of 斤 (catty) (533100). Similarly, the character 叹 (sigh) (3560977100) is generated by combining the radical code of 口 (mouth) with the phonetic code of 又 (again). The pattern in Fig. 5 is evident, with each row displaying Chinese characters sharing the same semantic component and radical code. Similarly, each column of Chinese characters represents the same (or similar) phonetic components with identical or similar phonetic codes.

The following list outlines the main steps of the image/text multimodal processing algorithm, with the complexity of $O(n^2)$, by combining SWPC with the image 0-1 matrix of Chinese characters:

Step 1: coding the root characters. The root characters are coded using the SWPC method and saved in a data set. Following the WB principle of character splitting, there are approximately 300 root characters. As an example, this section uses four root characters (口 (mouth), 女 (female), 斤 (catty), and 又 (again)), along with their corresponding radical and phonetic components, as well as the nine Chinese characters generated by them (Fig. 5). This step essentially labels the roots, radicals, and phonetic components of Chinese characters.

Step 2: pre-processing the images of the aforementioned root characters, including radicals and phonetic components. Each root (or radical or component) image is binarized using the Otsu threshold method (Otsu, 1979). This process converts the image into a matrix of zeros (0) and ones (1), where each matrix unit corresponds to a single image. This matrix is referred to as SWIC. Fig. 6 illustrates a 0-1 matrix of character 口 (mouth).

Step 3: establishing a Chinese character root image/text (coding) database. After digitizing all

Character root (字根编码)	Matrix (字/码矩阵)		Phonetic component (部件声码)			
			351100 口	951100 女	533100 斤	977100 又
口 mouth 35353535	R a d i c a l 偏旁 形码	口 3560	吕 name 3560351100		听 listen 3560533100	叹 sight 3560977100
女 woman 95959595		女 9575	如 as 9575351100	妁 quarrel 9575951100	妡 name 9575533100	奴 slave 9575977100
斤 catty 53515131		斤 5372			斫 catties 5372533100	
又 again 97979797		9774 又				双 double 9774977100

Fig. 5 Character matrix generated by the combination of radical and phonetic component codes

the images based on the coded root characters and their SWPC (from step 1), a databae of roots (as well as radicals and phonetic components) of Chinese characters is created. A simple root database is established with the four characters mentioned in step 1 (Fig. 5).

Step 4: similarity calculation between the image codes of Chinese characters (image matrix). We use the HM distance similarity (Hamming, 1950) to predict the similarity between image matrices of Chinese characters. The HM distance metric for binary inputs measures the number of differing positions between corresponding bits. In other words, it calculates the sum of differences between the bit units of two compared images. Therefore, the output of the HM function corresponds to the count of bit differences, where zero indicates complete image equality.

Step 5: multimodal processing of Chinese character using image/text. Based on the root database of characters (with the image matrix and SWPC of radical and component), various tasks of multimodal processing can be performed. Four roots and their associated radicals and components are processed to generate nine Chinese characters, which are listed in Fig. 5.

Step 6: result generation. Given a Chinese character, with the above steps our algorithm can identify which combinations of radical and phonetic components generate the similar character through similarity analysis. It is possible to provide the corresponding SWPC for understanding that character, as shown in Fig. 7.

After processing and comparing the image of a new character (source) with each image of the root

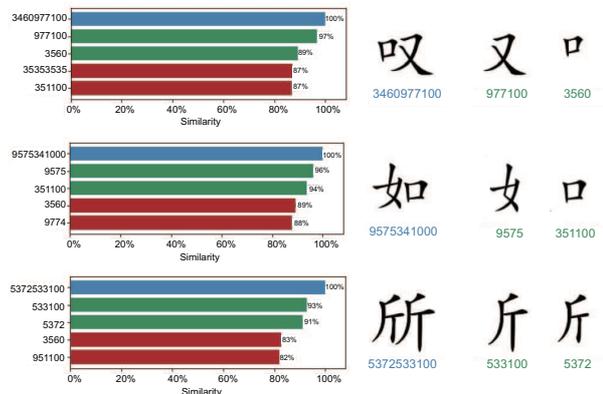**Fig. 6 Character 口 (mouth) by a 0-1 matrix**

Fig. 7 Correctly identified and generated Chinese characters using image and SWPC (References to color refer to the online version of this figure)

characters (target) in the root character database, the algorithm provides the similarity percentage, as shown in Fig. 7. This percentage represents the similarity between the source image and each target image. The chart in Fig. 7 represents this comparison and lists the three most similar symbols (including radicals or phonetic parts) between root characters and the image itself. Blue indicates the matching character (source and target), green indicates the correct prediction of radical and components, and red indicates the wrong prediction.

However, note that the proposed algorithm

using HM distance similarity is a simple method to solve this problem and has certain limitations. As mentioned above, it calculates the differences between image matrices, regardless of factors such as handwriting style, aesthetics, or change in radical positions (such as higher, lower, or inclined positions). Due to these limitations, some errors can occur, as seen in Fig. 8, where characters like 听 (listen) are incorrectly identified. This technical problem does not affect the basic principle of SWMP.
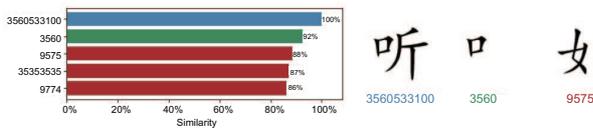

Fig. 8 Incorrectly identified Chinese characters using image and SWPC

Using the SWPC method, a root database with the image matrices and corresponding SWPC will be established. Because the WB input method can generate 300 roots, radicals, and components, the size of the database can still be managed. These components and their combinations can generate >10 000 Chinese characters. For any new Chinese character as the source, its similarity with each root in the database can be determined without the necessity to learn from any corpus, enabling image processing by "once learning" mechanism (Weigang and da Silva, 1999) to obtain the whole image of the character at one time.

Further research in this direction will be essential for advancing algorithms in pattern recognition, image synthesis, and generation. The use of soft shadow networks has great potential for a variety of applications (Sheng et al., 2023), including the generation of OOV characters without the need to consult any corpus or other sources.

7 SWPC for analogical reasoning

In this section, we establish analogical reasoning models that align with the morphological features of Chinese words and apply them to the Chinese analogical (CA8) data set and other Chinese idioms.

The CA8 data set was developed by the School of Chinese Information at Beijing Normal University (Li S et al., 2018). Its objective is to provide benchmark data and evaluate the analogical reasoning tasks in CNLP. The data set consists of two parts:

CA8-MOR-10177 comprises 10 177 pairs of questions concerning the formation of Chinese words, focusing on their morphological properties such as repetition, prefixes, and suffixes.

CA8-SEM-7636 contains 7636 pairs of questions related to the formation of Chinese words, highlighting the semantic rules of the Chinese language. The questions primarily cover 28 types of semantic relations from the domains of geography, history, nature, and human aspects.

7.1 Analogical reasoning method

Analogical reasoning, which explores language regularities, plays a crucial role in NLP. By calculating the representation vector of characters or words, the analogy problem can be solved given the word representations. For instance, examples like "apples–apple+car≈cars" illustrate morphological rules, while "king–man+woman≈queen" demonstrates semantic rules (Turney, 2012; Mikolov et al., 2013). Currently, many studies apply ML models, specifically word embedding techniques, to capture vector representations of words with semantic and syntactic properties. Pre-training on large corpora enables these models to enhance Chinese processing performance by providing insights into meaning and relationships.

Considering the pictophonetic features of Chinese characters and the Chinese language program, we divide the question pairs in the CA8 data set into two categories:

1. In standard language scenarios, new characters or words are generated based on Chinese language regulations and customs. If we have a question pair (A:B; C:X), where A, B, and C are known characters or words, and X is an unknown, we can use SWPC to encode the characters or words. Hereinafter, unless otherwise specified, $f(\cdot) = \text{SWPC}(\cdot)$. With $f(\text{A})$, $f(\text{B})$, and $f(\text{C})$, the generation of $f(X)$ of character or word, X, can be expressed as follows:

$$f(X) = f(\text{B}) - f(\text{A}) + f(\text{C}). \qquad (2)$$

The encoding requirement in Eq. (2) ensures that the differences between characters or words are distinguishable by morphological rules. In this case, the HM distance can be used to determine and characterize the similarities and differences between X and other characters.

2. In non-standardized language scenarios, new words are generated with maximum similarity. Different from the aforementioned scenario, it becomes necessary to calculate the similarity between characters or words due to the flexibility and idiomatic nature of Chinese. To account for uncertain language environments where exact but similar words may not exist, we can search for characters or words with the highest similarity to $f(X) =$ SWPC(X) relative to the known A, B, and C. This can be expressed as follows:

$$\text{Find } X: \arg\max\{\text{Sim}(f(X), f(B) - f(A) + f(C))\}. \tag{3}$$

Because this paper focuses on SWPC, our primary investigation revolves around the analogical reasoning problem that adheres to morphological regulations. Although Mikolov et al. (2013) demonstrated the effectiveness of cosine similarity in capturing syntactic and semantic relations, it necessitates further in-depth exploration, particularly for Chinese characters.

7.2 Analogical modes for CA8-Mor-10177

While Chinese character formation exhibits flexibility and diversity, it still adheres to fundamental rules, as explained in the classic Chinese grammar textbook (Wang L, 1959; Zhang B, 2008; Wang SK, 2016). Based on the CA8-Mor-10177 data subset (see Table 1 in Li S et al. (2018)), we present the basic pattern generation and similarity calculation methods for morphological characters using SWPC.

7.2.1 Repetitive character combination modes

The CA8-Mor-10177 data subset comprises various patterns involving character repetition, including A-character repetitive classes and three specific generation modes (A-A, A-一 (one)-A, A-来 (come)-A-去 (go)), totaling 2554 questions. There are also AB repetitive classes with three specific generation modes (A-A-B-B, A-里 (within)-A-B, A-B-A-B), amounting to 2535 questions. This subsection generalizes these patterns based on Chinese word formation methods and generates new words using SWPC.

1. (A, A-P-A) mode

The CA8-Mor-10177 data subset presents several (A, A-一 (one)-A) patterns, such as 避 (avoid) (A) and 补 (make up) (X): (避, 避一避; 补, 补一补).

Similar patterns exist in Chinese, such as 大 (large) (A) and 不 (no): (大, 大-不-大). Therefore, for this type of Chinese word formation, we propose a more general A-P-A model, which can be summarized as (A, A-P-A; B, X-P-X). The generation of a new word can be expressed as follows:

$$f(XPX) = \text{Concat}((f(AP) - f(A\mathbf{0}) + f(X\mathbf{0})), f(X)), \tag{4}$$

where Concat($\cdot$) is used to concatenate two strings of numbers, $\mathbf{0}$ is a vector with zeros, and its length is equal to that of the preceding word code, filling in the blanks. The correctness of Eq. (4) can be verified using the HM distance:

$$\begin{aligned}&\text{HMDis}(f(A), f(X))\\&= [\text{HMDis}(f(APA), f(XPX))]/2, \tag{5}\end{aligned}$$

where HMDis($\cdot$) represents the HM distance between the two strings of SWPC. To save space, the word generation modes described later exclude the verification modes based on the HM distance.

2. (A, A-P-A-Q) mode

The CA8-Mor-10177 data subset presents several questions: (说(A), 说(A)来(P)说(A)去(Q)); 比(X), 比(X)来(P)比(X)去(Q)), involving 说 (say), 来 (come), 去 (go), and 比 (compare). Chinese language exhibits similar patterns in other words, such as 能上能下 (being able to go up and down) and 不上不下 (not going up or down). Therefore, for these Chinese word patterns, we propose a more general A-P-A-Q model, namely (A, A-P-A-Q; X, X-P-X-Q). Using SWPC, we have

$$\begin{aligned}f(XPXQ) =&\text{Concat}((f(AP) - f(A\mathbf{0}) + f(X\mathbf{0})),\\&(f(AQ) - f(A\mathbf{0}) + f(X\mathbf{0}))), \tag{6}\end{aligned}$$

where the length of $\mathbf{0}$ equals that of the previous word code used to fill in the blanks.

3. (AB, A-P-AB) overlapping mode

The CA8-Mor-10177 data subset includes several questions, such as (慌张(AB), 慌(A), 里(P), 慌张(AB); 马虎(XY); 马(X), 里(P)马虎(XY)), involving 慌张 (panic), 里 (connection), and 马虎 (careless). This mode can be summarized as (AB, A-P-AB; XY, X-P-XY). Like the previous model, we have

$$\begin{aligned}f(XPXY) =&\text{Concat}((f(AP)\\&- f(A\mathbf{0}) + f(X\mathbf{0})), f(XY)). \tag{7}\end{aligned}$$

4. (AB, AA-BB) overlapping mode

Within the CA8-Mor-10177 data subset, there are several problems: (安全(AB), 安安(AA)全全(BB); 快乐(XY), 快快(XX)乐乐(YY)), involving 安全 (safety) and 快乐 (happy). The model can be summarized as follows: (AB, AA-BB; XY, XX-YY). Like the previous model, we have

$$f(XXYY) = f(AABB) - f(\mathbf{0}AB\mathbf{0})$$
$$+ f(\mathbf{0}XY\mathbf{0}) - f(A\mathbf{00}B) + f(X\mathbf{00}Y). \quad (8)$$

The data set encompasses not only the AB and ABAB overlapping patterns but also various other patterns commonly found in Chinese word formation, such as ABAC, ABCA, ABBC, ABCB, ABCC, and more. These patterns, including the AABB mode generated by SWPC, play a substantial role in Chinese character recognition and understanding the formation of Chinese words. While we do not delve into the specifics of these modes, it is important to acknowledge the significance of these patterns in CNLP.

7.2.2 Prefix word pattern

The CA8-Mor-10177 data subset includes 21 patterns of semi-prefixes, such as 大 (big), 小 (small), 老 (old), 第 (order), and 亚 (second), totaling 2553 questions. This pattern can be summarized as (A, PA; X, PX), e.g., 虎-老虎 (tiger-tiger) and 鹰-老鹰 (eagle-eagle). If A or P represents a multi-character word, the pattern can be extended to (AB, PQ-AB; XY, PQ-XY), e.g., 草原-大草原 (grassland-prairie) and 都市-大都市 (city-metropolis). Like the above formulas, we have the model as follows:

$$f(PQXY) = f(PQAB) - f(\mathbf{00}AB) + f(\mathbf{00}XY). \quad (9)$$

7.2.3 Suffix word pattern

The CA8-Mor-10177 data subset contains 41 patterns of semi-suffixes, including 者 (zhe), 式 (shi), 性 (sex), and others, totaling 2535 questions, e.g., 我 (I, A), 我们 (we, AP); 你 (you, X), 你们 (you, XP). This pattern can be summarized as (A, AP; X, XP). If A or P represents a multi-character word, the pattern can be extended to (AB, AB-PQ; XY, XY-PQ), e.g., 乐观, 乐观主义 (optimism-optimism) and 悲观, 悲观主义 (pessimism-pessimism). Referring to the above formulas, we have

$$f(XYPQ) = f(ABPQ) - f(AB\mathbf{00}) + f(XY\mathbf{00}). \quad (10)$$

7.3 Analogical modes from CA8-Sem-7636

Among the question pairs of the CA8-Sem-7636 data subset (Li S et al., 2018), 920 (12.05% of the total) adhere to the patterns of Chinese word formation discussed in Section 7.2. For instance, examples like (小学 (primary school), 小学生 (primary school student), 中学 (middle school), 中学生 (middle school student)) and (北京 (Beijing), 北京大学 (Peking University), 南京 (Nanjing), 南京大学 (Nanjing University)) fall into this category. When using SWPC, some question pairs change from following the semantic rule to follow the morphological rule, which can effectively and accurately generate new words for such problems.

There are 37 additional question pairs, which make up 0.48% of the data set. Due to the usage habits of the Chinese language, word formation in these cases does not strictly follow morphology. Examples include (葡萄藤grapevine, 葡萄grape; 芒果树mango tree, 芒果mango). Generating new words for these word pairs requires calculating the similarity between words using Eq. (3) with SWPC. For instance, as for {find word w: argmaxSim($f(w)$, f(葡萄grape) − f(葡萄藤grapevine) + f(芒果树mango tree))}, the cosine similarity to find 芒果 (mango) is 0.9463.

7.4 Comparative analysis with the baseline

Building upon the construction of the CA8 data set, Li S et al. (2018) conducted experimental analyses using the Skip-gram model with negative sampling (SGNS) (Mikolov et al., 2013) and positive point wise mutual information (PPMI) (Levy et al., 2015), along with three data preparation methods. The results can be found in Table 4 of Li S et al. (2018). For CA8-Mor-10177, using the SGNS model, the Chinese characters were added for word embedding, resulting in an accuracy of 0.455. For CA8-Sem-7636, the PPMI model and word+ngram were used for word embedding, resulting in an accuracy of 0.586. The findings here demonstrate that by analyzing the same data subset, 10 177 problems were predicted using SWPC, achieving 100% accuracy in generating new words. Analyzing the data subset of CA8-Sem-7636 led to the prediction of 920 questions, accounting for 12.05% of the data set, with 100% accuracy.

Table 4, which is reproduced from Kang et al.

(2019) and has been combined with our results, illustrates the comparison of performance for CA8-morphological data. These results highlight the applicability of SWPC in directly analyzing characters or words with standard language patterns for certain CNLP tasks.

Table 4 Comparison of the performance for CA8-morphological data

Model	Accuracy				
	A	AB	Prefix	Suffix	Subtotal
Skip-gram	0.023	0.089	0.003	0.252	0.111
CBOW	0.031	0.003	0.097	0.256	0.116
CWE	0.202	0.171	0.135	0.498	0.231
JWE	0.051	0.139	0.004	0.298	0.146
Cj2vec	0.188	0.173	0.181	0.521	0.289
SWPC (ours)	1.000	1.000	1.000	1.000	1.000

A and AB are categories in CA8-morphological data

7.5 Prompting by using SWPC

Note that the application of prompt engineering to language models, as discussed by Liu PF et al. (2023), highlights a new research direction. The concept of "pre-train, prompt, and predict" (PPP) holds significant potential for English-based language models and has been well received by the society. Simultaneously, the analogical reasoning model based on SWPC and Chinese grammar warrants in-depth discussion on how to conduct prompt engineering research in CNLP. If the SWPC analogical reasoning models proposed here and Chinese grammar can be integrated into Chinese language models, they have the potential to achieve 100% accuracy in prompt-based learning by question answering. This would allow us to fine-tune Chinese LLMs and achieve a better PPP effect for thousands of Chinese idioms and other related CNLP tasks.

We devised a Chinese Q&A prompt using the CA8-SEM-7636 data set (Li S et al., 2018) to assess the performance of LLMs, such as ChatGPT and Google Bard: Please use the following forms: (1) (XP, YP; XQ, YQ): (公狼male wolf, 母狼female wolf; 公熊male bear, 母熊female bear); (2) (XP, YP; MR, FR): (公狼male wolf, 母狼female wolf; 雄鸟male bird, 雌鸟female bird); (3) to have the correct animal name (XQ, YQ; MT, ?): (公熊, 母熊; 雄兔male rabbit, ?).

Initially, ChatGPT provided an incorrect re-

sponse. However, upon prompt modification and training with the accurate answer, ChatGPT subsequently produced the correct response. In contrast, Google Bard initially generated an erroneous answer. Despite our efforts to rectify this through training, Google Bard continued to produce inaccurate responses. This highlights the need for substantial improvement in some LLMs' ability to generate coherent Chinese text.

Liu PF et al. (2023) proposed Eq. (11) to search over the set of potential answers "z" by calculating the probability of their corresponding filled prompts using a pre-trained language model $P(.;\theta)$:

$$\hat{z} = \text{search } P(f_{\text{fill}}(x', z); \theta), \quad z \in \mathbb{Z}, \qquad (11)$$

where θ is the learning parameter from LLMs. Using word embedding to represent the characters or words (Mikolov et al., 2013), as in Eq. (3), we have

$$\text{Find } z : \text{argmaxSim}(f(z), f(\text{MR}) - f(\text{MQ}) + f(\text{FQ})), \qquad (12)$$

where $f(\cdot) = \text{WE}(\cdot)$, and $\text{WE}(\cdot)$ is the word embedding code using ML. Some experimental performances are reported for this example in Table 4; the accuracy is no more than 58.6% (Li S et al., 2018) and ≤42.33% (Jin et al., 2022).

We have introduced a range of analogical reasoning models using SWPC. These models can be effectively applied to the prompting functions put forth by Liu PF et al. (2023). For an illustration, Table 5 shows the analogical reasoning using the prompt method for the above question.

All the Chinese characters (words) in Table 5 are represented by SWPC, which is used for the prompting and prediction processes. In this example, the Chinese program and habits are used to ensure reasoning (prompting) processing. In some cases, it is necessary to calculate the similarity. Similar to Eq. (12), when applying the SWPC approach, $f(\cdot) = \text{SWPC}(\cdot)$. There is $z = \text{FR}$; i.e., in our example, F=雌 (female), R=兔 (rabbit).

In the Chinese language, thousands of words are formed by idioms (成语) and proverb (习语) according to Chinese program and habits. These words can be transferred to (AP, BP; MQ, FQ) and related modes; see Eqs. (4)–(10) and also refer to the CA8 data set (Li S et al., 2018). The search probability $P(\cdot; \theta)$ in Eq. (11) is 1.00; in other words, the accuracy in Eq. (12) reaches 100%. In this case, we may

Table 5 Solving the analogical reasoning problem by prompt mode*

Name	Notation	Example	Description
Input	x	(公熊, 母熊; 雄鸟, 雌鸟) ->	One or multiple texts
Output	y	(公熊, 母熊; 雄兔, ?)	Output other text
Prompting function	$f_{\mathrm{prompt}}(x)$	(AP, BP; MQ, FQ) ->(AP, BP; [X], [Z})	A function that converts the input into a specific form by inserting the input x and adding a slot $[Z]$, where z will be filled later
Prompt	x'	(AP, BP; MQ, FQ) ->(AP, BP; MR, [Z})	$[X]$ is instantiated by input $x' = $ MR but answer slot $[Z]$ is not
Fill prompt	$f_{\mathrm{fill}}(x',z)$	(MQ, FQ) ->(MR, ?R)	A prompt where slot $[Z]$ is filled with any character related to R
Answered prompt	$f_{\mathrm{fill}}(x',z*)$	(MQ, FQ) ->(MR, FR)	A prompt where slot $[Z]$ is filled with a true answer FR
Answer	z	FR (雌兔), XR (母兔), ...	To verify the Chinese program and customs

* Using a form similar to that in Liu PF et al. (2023)

regard the Chinese language as a special LLM, which can be more effectively used for analogical reasoning and prompting.

The Chinese language possesses some characteristics that give rise to special grammatical rules and habits in character/word formation. Currently, several LLMs do not adequately address these intricacies in Chinese. However, SWMP/SWPC methods try to effectively capture these distinctive features and significantly boost the development of Chinese LLMs, yielding twice the results with half the effort.

8 Fine-tuning of similarity by SWPC

In 2019, the Department of Computer Science and Technology at Tsinghua University developed the COS960 test data set (Huang JJ et al., 2019), which studies the similarity between word embedding models and Chinese characters from related corpora. The data set consists of 960 pairs of words, and the similarity score specified manually is used as the benchmark. The question pairs were designed in the format of (唯独, 惟独only 3.7333), where 3.7333/4 represents the average similarity based on 15 manual evaluations. We discuss the calculation and comparison of fine-tuning the similarity using the FC number, WB code, and SWPC.

8.1 Similarity between Chinese word pairs

SWPC is primarily used to calculate the similarity of the word pairs and analyze the effectiveness of the proposed method. Similarity calculation methods include cosine similarity and HM distance

similarity, facilitating analysis and comparison. The following presents the similarity calculation of several word pairs in different language scenarios.

When the semantics and forms of Chinese characters are similar, and the number of paired words is the same or slightly different, the similarity between the word pairs can be calculated directly using modes such as (AC ≈ BC) and (AB ≈ AC).

For instance, SimCos(f(唯独unique), f(惟独unique)) = 3.8084, the average score of human evaluation is 3.7333, and the relative error is 2.01%. SimCos(f(火爆fire storm), f(火暴fire storm)) = 3.7870, the average score of human evaluation is 3.6667, and the relative error is 3.28%.

When the semantics and forms of Chinese characters are similar, the number of paired words is the same or slightly different, but the positions are different. Most of these word pairs are Chinese words with different orders (AB = BA). The similarity between two words can be calculated using the following equation:

$$\mathrm{Sim}(f(\mathrm{AB}),(\mathrm{BA})) = \max(\mathrm{Sim}(f(\mathrm{AB}),f(\mathrm{BA})), \\ \mathrm{Sim}(f(\mathrm{AB}),f(\mathrm{AB}))), \quad (13)$$

where $f(\cdot) = \mathrm{SWPC}(\cdot)$. For word pair (躲闪, 闪躲dodge), there is Sim(f(躲闪), f(闪躲)) = max(Sim(f(躲闪), f(闪躲)), Sim(f(躲闪), f(躲闪)))=1.00, which is 3.45% higher than the manual score 3.8667. There will also be some (AC ≈ CB) modes, e.g., (帷幔, 幔帐curtain). There is Sim(f(帷幔), f(幔帐)) = max(HamSim(f(帷幔), f(幔帐)), HamSim(f(帷幔), f(帐幔)), HamSim(f(幔帷), f(幔帐))) = max(3.6480, 3.8860, 3.8860) = 3.8860, which is 29.53% higher than the manual score 3.0.

The question regarding the customary formation of Chinese words remains. For instance, it is uncertain whether the combination of Chinese characters CA or BC is in accordance with the conventions of Chinese words. These questions will be explored in the study of Six-Writings semantic coding (SWSC).

8.2 Analysis of similarity of the COS960 dataset

Table 6 presents a comparative analysis of the similarity calculation for 960 word pairs in the COS960 test data set. The first column of the table represents the manual scores assigned to the data set, categorized into six intervals serving as benchmark data. The subsequent columns display the relative errors between the benchmark scores and the similarities computed using FC, WB, and SWPC, separately. The similarity calculation methods include cosine similarity and HM similarity.

Table 6 Relative error of similarity of COS960 data

Interval	Relative error					
	FC		WB		SWPC	
	COS	HM	COS	HM	COS	HM*
Sim=4	0.186	0.400	0.113	0.375	0.147	0.388
3≤Sim<4	0.116	0.317	0.086	0.265	0.099	0.291
2≤Sim<3	0.146	0.167	0.205	0.154	0.175	0.161
1≤Sim<2	0.425	0.217	0.477	0.212	0.425	0.214
0≤Sim<1	0.511	0.117	0.665	0.150	0.607	0.133
Sim=0	0.548	0.200	0.729	0.188	0.653	0.194
Mean	0.339	0.200	0.461	0.192	0.404	0.196

HM* denotes the average value of the HM similarity index calculated using both FC numbers and SWPC, while Mean represents the overall average across all 960 word pairs

The analysis results of similarity calculation for the whole COS960 data set are listed below:

1. For word pairs with high similarity (2≤Sim≤4), three codes (FC, WB, and SWPC) are used, and the cosine similarity shows good agreement with the benchmark. However, when dealing with word pairs with low similarity (0≤Sim<2), some dissimilar word pairs result in higher similarity scores. The results in Table 6 do not accurately reflect their true similarity.

2. When considering word pairs with high similarity (3≤Sim≤4), using the three codes, the calculation error of HM similarity is large, with an average relative error above 26%. However, for word pairs with low similarity (0≤Sim<3), the calculation results of HM similarity are better, with an average relative error of less than 20%. Even for completely dissimilar word pairs, the dissimilarity between them can be clearly distinguished.

3. The HM similarity is highly sensitive to the pictophonetic coding of Chinese characters. Particularly, in the first 26 word pairs with high similarity (Sim=4) in the COS960 test data set, four pairs of Chinese words exhibit the property of different order (AB ≈ BA). If this factor is not considered, the average relative error of HM approximation for these 26 words is 44.21%. However, if this factor is taken into account and the similarity calculation is corrected, the average relative error of HM approximation is reduced to 32.80%. Table 6 demonstrates that HM similarity calculation can effectively reflect the similarity and dissimilarity between Chinese word pairs.

4. The calculation results of FC, WB, and SWPC in the two similarity methods are relatively stable, aligning with the analysis results mentioned earlier. Considering the advantages of SWPC's combination of FC and WB coding, the results presented in the HM* column are more robust and credible. Particularly, the calculation of HM similarity can be used to fine-tune the results of CNLP tasks, such as word embedding, allowing for the reflection of nuanced aspects of Chinese characters.

9 Conclusions and future work

With the rapid advancement of AI, CNLP has made significant progress. However, the accuracy of certain tasks in CNLP requires further improvement. We propose the SWMP framework for Chinese language models. This framework integrates multimodal information of Chinese characters, including pictophonetics, pinyin, images, and semantics, to enhance the effectiveness of CNLP.

By conducting variability analysis of the pictophonetic coding of Chinese characters, such as FC numbers, and exploring similarity calculation methods, we propose to augment and normalize the WB numerical coding from 11–55 to 11–99. At the same time, the numerical coding method of pinyin is developed; i.e., the initial consonants are coded by the GB uppercase letters/numbers and the vowels are coded by GB lowercase letters/numbers. Then they are augmented and normalized in the range of 11–99.

Under the unified coding framework of SWMP, we introduce the concept of SWPC, which combines the expression of characters with Chinese grammar and flexible properties. With its moderate granularity of representation, SWPC possesses a generative and prompting mechanism for multimodal processing of Chinese characters and graphics, complementing existing language models using word embedding methods. The specific applications of SWPC, and also our contributions, are as follows:

1. Considering the variability of numerical expressions of Chinese characters and the analysis of related similarity calculation methods (Wang JT, 2011), SWPC can effectively express the similarity between similar characters and the dissimilarity between different characters.

2. By combining SWPC with Chinese character image processing and other multimodal processing technologies, we propose different methods for Chinese character generation. Through simple examples, our research demonstrates how to establish a Chinese character data set including roots, radicals, and phonetic components, based on their SWPC, thus forming a Chinese character generative matrix that facilitates various CNLP tasks.

3. By leveraging Chinese grammar, the phonological features of Chinese characters, and the requirements of language models, we establish analogical reasoning models for various word combinations using SWPC. As a result, the processing accuracy of word pairs, including repetition (AABB), prefix (PQAB), and suffix (ABPQ), can achieve 100% accuracy. These models are also used for the purpose of prompting Q&A Chinese language models.

4. The application of SWPC to data sets like CA8 (Li S et al., 2018) enables high-precision analogical reasoning for word pairs conforming to Chinese grammar and pictophonetic properties. Given that 90% of Chinese characters exhibit phono-semantic characteristics, the proposed method holds considerable potential for practical applications.

5. SWPC is applied in analyzing the COS960 data set (Huang JJ et al., 2019) to evaluate the strengths and weaknesses of various similarity calculation methods. The objective is to improve the prediction accuracy of word similarity. At the same time, the SWPC approach can be used as a complementary technology to fine-tune word embedding results based on ML.

The core and essence of the ancient Six-Writings lie in "explaining words by writing" (说文解字) in multimodal processing. Our proposed SWMP method is not limited to CNLP (including simplified and traditional Chinese characters), but can be applied to other non-alphabetic languages such as Japanese, Korean, and Vietnamese. This comprehensive approach aims to enhance the effectiveness of applications in a multi-lingual environment.

The SWMP framework we propose is just the first step in improving Chinese language models using the concept of Six-Writings. It still has some shortcomings that need to be addressed. For example, to enable multimodal processing of pictophonetic, pinyin, image, property, audio/video, and understanding, we need to integrate SWMP into the language model, or even establish a new Chinese language model. This integration should also involve the fusion of these various types of information.

Furthermore, we have presented only the theory and methodology of SWMP/SWPC and not addressed the establishment of a Chinese character database (e.g., 3500 common Chinese characters), which would facilitate SWPC/image coding based on roots, radicals, and components. It is necessary to establish a common Chinese character root database to realize the coding of roots, radicals, and components by SWPC. However, the number of characters, the granularity of character splitting, and the standardization of coding all require the support from the academic community and even the government. It is also necessary to develop a more appropriate coding/image multimodal analysis ML algorithm using SWPC associated to pinyin/image/audio/video, and to gradually introduce SWPC into Chinese LLMs and related tasks. Finally, we should strengthen the coding of semantic features of the Chinese words, combine it with word embedding technology, and conduct application research in the CNLP field.

Contributors

Acknowledgements

This article pays tribute to Shen XU (许慎) and his monumental work, *Shuo Wen Jie Zi* (说文解字). The authors extend their gratitude to the inventors of the four-corner number, Cangjie, Wubi, Zheng code, and others. The authors are also very grateful for the valuable comments and suggestions of anonymous reviewers.

Compliance with ethics guidelines

Li WEIGANG is an editorial board member of *Frontiers of Information Technology & Electronic Engineering*, and he was not involved with the peer review process of this paper. All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Cao SS, Lu W, Zhou J, et al., 2017. Investigating stroke-level information for learning Chinese word embeddings. Proc 16th Int Semantic Web Conf.

Cao SS, Lu W, Zhou J, et al., 2018. cw2vec: learning Chinese word embeddings with stroke n-gram information. Proc 32nd AAAI Conf on Artificial Intelligence, 30th Innovative Applications of Artificial Intelligence Conf, and 8th AAAI Symp on Educational Advances in Artificial Intelligence, p.5053-5061.

Chen HY, Yu SH, Lin SD, 2020. Glyph2Vec: learning Chinese out-of-vocabulary word embedding from glyphs. Proc 58th Annual Meeting of the Association for Computational Linguistics, p.2865-2871.
https://doi.org/10.18653/v1/2020.acl-main.256

Chen XX, Xu L, Liu ZY, et al., 2015. Joint learning of character and word embeddings. Proc 24th Int Conf on Artificial Intelligence, p.1236-1242.

Everitt BS, Skrondal A, 2010. The Cambridge Dictionary of Statistics (4th Ed.). Cambridge University Press, Cambridge, UK.

Feng ZW, 2012. A Concise Course of Natural Language Processing. Shanghai Foreign Language Education Press, Shanghai, China (in Chinese).

Gao P, 2003. Standard Tutorial of Wubi Font Input Method. Science Press, Beijing, China (in Chinese).

Hamming RW, 1950. Error detecting and error correcting codes. *Bell Syst Tech J*, 29(2):147-160.
https://doi.org/10.1002/j.1538-7305.1950.tb00463.x

Huang BR, Li W, 2012. Contemporary Chinese Language. Peking University Press, Beijing, China (in Chinese).

Huang JJ, Qi FC, Yang CH, et al., 2019. COS960: a Chinese word similarity dataset of 960 word pairs.
https://arxiv.org/abs/1906.00247

Jin H, Zhang ZB, Yuan PP, 2022. Improving Chinese word representation using four corners features. *IEEE Trans Big Data*, 8(4):982-993.
https://doi.org/10.1109/TBDATA.2021.3106582

Kang RZ, Zhang HJ, Hao WN, et al., 2019. Learning Chinese word embeddings with words and subcharacter n-grams. *IEEE Access*, 7:42987-42992.
https://doi.org/10.1109/ACCESS.2019.2908014

Levy O, Goldberg Y, Dagan I, 2015. Improving distributional similarity with lessons learned from word embeddings. *Trans Assoc Comput Ling*, 3:211-225.
https://doi.org/10.1162/tacl-a-00134

Li BA, Li Y, Meng QC, 2005. Chinese Information Processing Technology: Principles and Applications. Tsinghua University Press, Beijing, China (in Chinese).

Li S, Zhao Z, Hu RF, et al., 2018. Analogical reasoning on Chinese morphological and semantic relations. Proc 56th Annual Meeting of the Association for Computational Linguistics, p.138-143.
https://doi.org/10.18653/v1/P18-2023

Liu MD, Liang X, 2021. A method of Chinese character glyph similarity calculation based on radical knowledge representation learning. *J Chin Inform Process*, 35(12):47-59 (in Chinese).
https://doi.org/10.3969/j.issn.1003-0077.2021.12.005

Liu PF, Yuan WZ, Fu JL, et al., 2023. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing. *ACM Comput Surv*, 55(9):195. https://doi.org/10.1145/3560815

Lu W, Zhang ZB, Yuan PP, et al., 2022. Learning Chinese word embeddings by discovering inherent semantic relevance in sub-characters. Proc 31st ACM Int Conf on Information & Knowledge Management, p.1369-1378.
https://doi.org/10.1145/3511808.3557376

Meng YX, Wu W, Wang F, et al., 2019. Glyce: Glyph-vectors for Chinese character representations. Proc 33rd Int Conf on Neural Information Processing Systems, p.2742-2753.

Mikolov T, Yih WT, Zweig G, 2013. Linguistic regularities in continuous space word representations. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.746-751.

Otsu N, 1979. A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern*, 9(1):62-66.
https://doi.org/10.1109/TSMC.1979.4310076

Petrov A, la Malfa E, Torr PH, et al., 2023. Language model tokenizers introduce unfairness between languages.
https://arxiv.org/abs/2305.15425

Saleh AA, Weigang L, 2023. Deep self-organizing cube: a novel multi-dimensional classifier for multiple output learning. *Expert Syst Appl*, 230:120627.
https://doi.org/10.1016/j.eswa.2023.120627

Schulman J, Zoph B, Kim C, 2022. Introducing ChatGPT. https://openaicom/blog/chatgpt [Accessed on May 30, 2023].

Sheng YC, Zhang JM, Benes B, 2021. SSN: soft shadow network for image compositing. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.4378-4388. https://doi.org/10.1109/CVPR46437.2021.00436

Sheng YC, Liu YF, Zhang JM, et al., 2022. Controllable shadow generation using pixel height maps. 17th European Conf on Computer Vision, p.240-256.
https://doi.org/10.1007/978-3-031-20050-2_15

Sheng YC, Zhang JM, Philip J, et al., 2023. PixHt-Lab: pixel height based light effect generation for image compositing. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.16643-16653.
https://doi.org/10.1109/CVPR52729.2023.01597

Song JH, Li GY, Wang N, 2006. Productive representation on the phonetic-semantic relations of *Shuowenjiezi*. *J Chin Inform Process*, 20(2):53-59 (in Chinese). https://doi.org/10.3969/j.issn.1003-0077.2006.02.008

Standardization Administration of the People's Republic of China, 2022. Information Technology - Chinese Coded Character Set. GB 18030-2022. National Standards of People's Republic of China (in Chinese).

Su TR, Lee HY, 2017. Learning Chinese word representations from glyphs of characters. Proc Conf on Empirical Methods in Natural Language Processing, p.264-273. https://doi.org/10.18653/v1/D17-1025

The Unicode Consortium, 2022. The Unicode Standard, Version 15.00. The Unicode Consortium. Mountain View, CA, USA.

The Wubi Group, 2000. Wubi code: a method for inputting Chinese characters. *Chin J Inform Process*, 24(3):1-10 (in Chinese).

Turney PD, 2012. Domain and function: a dual-space model of semantic relations and compositions. *J Artif Intell Res*, 44(1):533-585. https://doi.org/10.5555/2387933.2387945

Wang JT, 2011. Research towards Chinese string similarity based on the clustering feature of Chinese characters. *New Technol Lib Inform Ser*, (2):48-53 (in Chinese).

Wang L, 1959. Chinese Modern Grammar. Zhonghua Book Company, Hong Kong, China (in Chinese).

Wang SK, 2016. New Modern Chinese Course. Shanghai Jiao Tong University Press, Shanghai, China (in Chinese).

Wang SR, Zhou W, Zhou Q, 2020. Radical and stroke-enhanced Chinese word embeddings based on neural networks. *Neur Process Lett*, 52(2):1109-1121. https://doi.org/10.1007/s11063-020-10289-6

Weigang L, da Silva NC, 1999. A study of parallel neural networks. Proc Int Joint Conf on Neural Networks, p.1113-1116. https://doi.org/10.1109/IJCNN.1999.831112

Weigang L, Enamoto LM, Li DL, et al., 2022. New directions for artificial intelligence: human, machine, biological, and quantum intelligence. *Front Inform Technol Electron Eng*, 23(6):984-990. https://doi.org/10.1631/FITEE.2100227

Xu J, Liu JW, Zhang LG, et al., 2016. Improve Chinese word embeddings by exploiting internal structure. Proc Conf of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, p.1041-1050. https://doi.org/10.18653/v1/N16-1119

Xu S, 1997. Discussing Writing and Explaining Characters. Yuelu Publishing House, Changsha, China (in Chinese).

Yeromiyan T, 2022. The Six Types of Chinese Characters. https://studycli.org/chinese-characters/types-of-chinese-characters/ [Accessed on May 30, 2023].

Yu JX, Jian X, Xin H, et al., 2017. Joint embeddings of Chinese words, characters, and fine-grained subcharacter components. Proc Conf on Empirical Methods in Natural Language Processing, p.286-291. https://doi.org/10.18653/v1/D17-1027

Zhang B, 2008. Newly Edited Chinese Language (2nd Ed.). Fudan University Publishing, Shanghai, China (in Chinese).

Zhang Y, Liu YG, Zhu JJ, et al., 2019. Learning Chinese word embeddings from stroke, structure and pinyin of characters. Proc 28th ACM Int Conf on Information and Knowledge Management, p.1011-1020. https://doi.org/10.1145/3357384.3358005

Zhang ZB, Zhong ZM, Yuan PP, et al., 2023. Improving entity linking in Chinese domain by sense embedding based on graph clustering. *J Comput Sci Technol*, 38(1):196-210. https://doi.org/10.1007/s11390-023-2835-4

Zhao DP, Xiong HX, Tian FS, et al., 2021. Research on Chinese text similarity calculation based on sequence alignment algorithm. *Lib Inform Serv*, 65(11):101-112 (in Chinese). https://doi.org/10.13266/j.issn.0252-3116.2021.11.011

Zhao YR, 2017. A Grammar of Spoken Chinese. University of California Press, CA, USA.

Zhou J, Ke P, Qiu XP, et al., 2023. ChatGPT: potential, prospects, and limitations. *Front Inform Technol Electron Eng*, early access. https://doi.org/10.1631/FITEE.2300089

Zhou JN, Wang JK, Liu GS, 2019. Multiple character embeddings for Chinese word segmentation. Proc 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, p.210-216. https://doi.org/10.18653/v1/P19-2029

Zhuang CY, Zheng YJ, Huang WH, et al., 2019. Joint fine-grained components continuously enhance Chinese word embeddings. *IEEE Access*, 7:174699-174708. https://doi.org/10.1109/ACCESS.2019.2956822