Scientific Research Publishing

Journal of Data Analysis and Information Processing, 2023, 11, 200-216
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Typos Correction in Overseas Chinese Learning Based on Chinese Character Semantic Knowledge Graph

Jing Xiong[1,2*], Xue Zhai[3], Zhan Zhang[1,2], Feng Gao[1,2]

[1]School of Computer and Information Engineering, Anyang Normal University, Anyang, China
[2]Key Laboratory of Oracle Bone Inscriptions Information Processing, Ministry of Education, Anyang, China
[3]School of Civil Engineering and Architecture, Anyang Normal University, Anyang, China
Email: *jingxiong125@gmail.com

How to cite this paper: Xiong, J., Zhai, X., Zhang, Z. and Gao, F. (2023) Typos Correction in Overseas Chinese Learning Based on Chinese Character Semantic Knowledge Graph. *Journal of Data Analysis and Information Processing*, **11**, 200-216.
https://doi.org/10.4236/jdaip.2023.112011

Received: April 26, 2023
Accepted: May 28, 2023
Published: May 31, 2023

Abstract

In recent years, more and more foreigners begin to learn Chinese characters, but they often make typos when using Chinese. The fundamental reason is that they mainly learn Chinese characters from the glyph and pronunciation, but do not master the semantics of Chinese characters. If they can understand the meaning of Chinese characters and form knowledge groups of the characters with relevant meanings, it can effectively improve learning efficiency. We achieve this goal by building a Chinese character semantic knowledge graph (CCSKG). In the process of building the knowledge graph, the semantic computing capacity of HowNet was utilized, and 104,187 associated edges were finally established for 6752 Chinese characters. Thanks to the development of deep learning, OpenHowNet releases the core data of HowNet and provides useful APIs for calculating the similarity between two words based on sememes. Therefore our method combines the advantages of data-driven and knowledge-driven. The proposed method treats Chinese sentences as subgraphs of the CCSKG and uses graph algorithms to correct Chinese typos and achieve good results. The experimental results show that compared with keras-bert and pycorrector + ernie, our method reduces the false acceptance rate by 38.28% and improves the recall rate by 40.91% in the field of learning Chinese as a foreign language. The CCSKG can help to promote Chinese overseas communication and international education.

Keywords

Chinese Character Meaning, Knowledge Graph, Typos Correction, OpenHowNet, Semantic Relevancy

1. Introduction

Chinese characters have been used continuously for the longest time so far. They are the only characters in the major writing systems of ancient times that have been passed down till now. Other ancient characters such as Hieroglyphics and Cuneiform have disappeared. Only Chinese characters still have a lasting vitality. They are also the only highly developed ideographic characters that are still widely used in the world. Ideograph is a system of writing words or morphemes in symbolic writing, which does not directly or simply represent speech.

Chinese characters are among the most widely adopted writing systems in the world by the number of users. In recent years more and more foreign friends begin to learn Chinese characters. The majority of Chinese characters are pictophonetic characters, accounting for 85% of the total. A pictophonetic character is composed of two parts: the descriptive component and the phonetic component. The descriptive component usually gives the hint of the boundary of the character's meaning, while the phonetic one suggests the pronunciation of the character. Therefore, many foreigners learn Chinese characters employing the descriptive component or the phonetic component. However, due to the development of the glyphs and the phonetic system of Chinese characters, many Chinese characters have the phenomenon that the descriptive component does not indicate the meaning or the phonetic component does not indicate the pronunciation. Even for native Chinese speakers, the pronunciation of Chinese characters cannot be simply deduced from their glyphs, and the specific meanings of Chinese characters cannot all be deduced from their descriptive components. Not to mention foreign Chinese beginners whose native language is phonetic [1]. It is not always possible to get consistent results according to the structure of the six categories of Chinese characters, because the same character may be classified into different categories by different experts [2]. The meaning expressed by Chinese characters is related to the ideograph used, while the meaning of most Chinese characters is the extension of the concept expressed by ideograph, or the expansion or reduction of the meaning of words. Therefore, if the knowledge structure of ideograph is mastered, it is equivalent to mastering the meaning of most Chinese characters. The Chinese character writing system represents and classifies lexical units according to semantic classes. So it is important to learn the meaning of the Chinese character glyph.

Being proficient in Chinese is difficult, and even native Chinese speakers often make mistakes. After reviewing about 3000 books, 1000 journals, and 100 newspapers, the Journal of *YAOWEN JIAOZI* has sorted out a batch of mistake characters based on the frequency of errors and experts' comments [3]. The top 10 of them are shown in **Table 1**.

In Chinese communication and international education, the understanding of Chinese characters is the most basic requirement. However, the current Chinese character learning methods, beginners mainly rely on the Chinese character glyph to learn, it often encounters great obstacles, such as UTF8gbsn "肓(the organ between the heart and the diaphragm)" and "盲(blind)", "粟(millet)" and

Table 1. The top 10 mistake characters of native Chinese speakers (the characters in the brackets are correct, and the letters in square brackets are the Pinyin of Chinese characters).

No.	mistake characters	No.	mistake characters
1	松驰[chí] (弛[chí])	6	挖墙角[jiǎo] (脚[jiǎo])
2	穿[chuān] (川[chuān]) 流不息	7	再接再励[lì] (厉[lì])
3	渡[dù] (度[dù])假村	8	谈笑风声[shēng] (生[shēng])
4	一幅[fú] (副[fù])对联	9	渲[xuàn] (宣[xuān])泄
5	既[jì] (即[jí])使	10	九洲[zhōu] (州[zhōu])

"粟(chestnut)", it difficult to distinguish and understand them through the glyph. From the HSK Dynamic Composition Corpus [4], we have calculated the top 10 Chinese characters (the group marked in blue) that foreign learners are most likely to make mistakes, as shown in Figure 1.

In Figure 1, the common Chinese character mistakes come from the HSK Dynamic Composition Corpus. Two groups are compared in the figure, where the blue one is the common mistakes of foreign learners and the red one is the common mistakes of native Chinese speakers. And those characters in the red group are corresponding to Table 1.

Comparing Figure 1 and Table 1, we can find that Chinese beginners and native speakers have very different levels of understanding of Chinese. The top 10 most common wrong Chinese characters listed in Table 1 do not maintain the same trend in the HSK Dynamic Composition Corpus. Therefore, it is inconsistent that foreigners who want to learn Chinese use the same method to study Chinese. We should base on the actual situation and needs of foreign learners and provide targeted learning methods.

The mistake Chinese characters can be divided into wrongly written characters and mispronounced characters. With the help of the Chinese input methods, there is little chance of wrongly writing, so most mistakes belong to mispronounced characters. The common causes of mispronounced characters are similarity in the glyph, similarity in pronunciation, similarity in meaning, similarity in both glyph and pronunciation, similarity in glyph, pronunciation and meaning. However, correct recognition of mispronounced characters requires an accurate understanding of the meaning of Chinese characters.

For example, several common mistakes are shown in Table 2. They come from the real composition exams corpus [4] of foreign students who are learning Chinese. It is worth noting that these Chinese characters are all sorted out from the handwritten answer sheets.

In Table 2, the left side of "|" is English, and the right side of "|" is Chinese. As can be seen from Table 2, it is not enough to learn Chinese characters only by their glyphs and pronunciations. It is more important to learn and compare them from a semantic point of view. If Chinese character learning is integrated into the understanding of the semantics, the learning efficiency can be greatly improved. For example, the character "片(slice)" is related to "木(wood)" and it means to split "木(wood)" into two halves [5]. Their semantic correlation is shown in Figure 2.

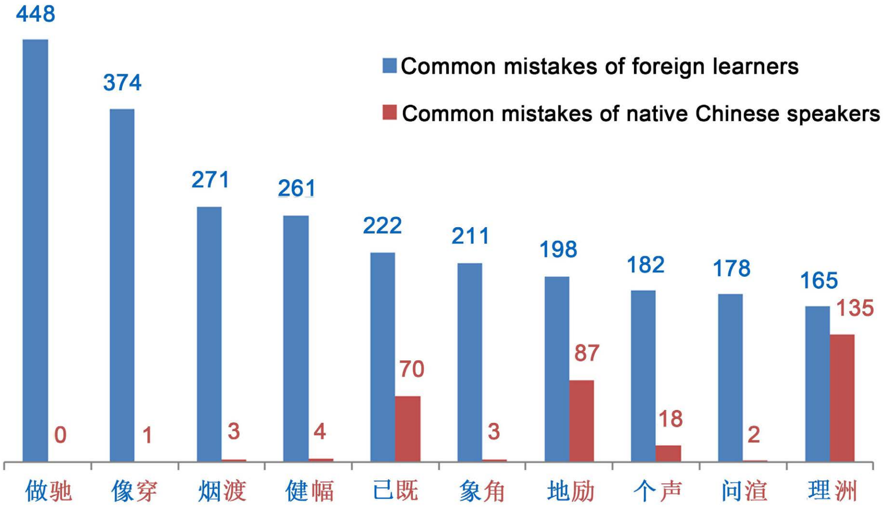


Figure 1. The top 10 frequent Chinese characters of common mistakes.

Table 2. Samples of common mistakes in Chinese learning.

Correct expression	Wrong expression	Error causes
consider | 考[kǎo]虑[lǜ]	考[kǎo]虚[xū]，考[kǎo]虎[hǔ]	Mistakes due to similar glyphs, misunderstanding of meaning
consider |考[kǎo]虑[lǜ]	考[kǎo]滤[lǜ]	Mistakes due to similar glyphs and the same pronunciation, misunderstanding of meaning
impression |印[yìn]象[xiàng]	印[yìn]像[xiàng]，印[yìn]相[xiàng]，印[yìn]响[xiǎng]，影[yǐng]像[xiàng]，影[yǐng]想[xiǎng]，影[yǐng]响[xiǎng]，影[yǐng]象[xiàng]，应[yǐng]向[xiàng]，应[yìng]像[xiàng]	Mistakes due to similar glyphs or the same or similar pronunciation, misunderstanding of meaning
impression |印[yìn]象[xiàng]	印[yìn]影[yǐng]，形[xíng]相[xiàng]	Misunderstanding of meaning
imagine |想[xiìng]象[xiàng]	想[xiǎng]像[xiàng]，想[xiǎng]相[xiàng]，想[xiǎng]想[xiǎng]，想[xiǎng]向[xiàng]	Mistakes due to similar glyphs or the same or similar pronunciation
already | 已[yǐ] 经[jīng]	己[jǐ]经[jīng]，巳[sì]经[jīng]，之[zhī]经[jīng]，巴[bā]经[jīng]	Mistakes due to similar glyphs
already |已[yǐ] 经[jīng]	以[yǐ_]经[jīng]，乙[yǐ]经[jīng]，一[yī]经[jīng]	Mistakes due to similar glyphs or the same or similar pronunciation

It can be seen from **Figure 2** that the correlation between “木(wood)” and “片(slice)” is established through “cut”. Taking “木(wood)” and “片(slice)” as nodes and “cut” as edges to form a semantic association diagram, then “木(wood)” and “片(slice)” can be combined into a knowledge unit to establish a common cognitive model, greatly improving the traditional learning mode of Chinese characters.

All the Chinese characters have such semantic association relations. If the Chinese characters are associated together using their semantics, and thus form a large knowledge graph, the Chinese learners can use it to master a large number of Chinese characters quickly. For the construction of the Chinese character

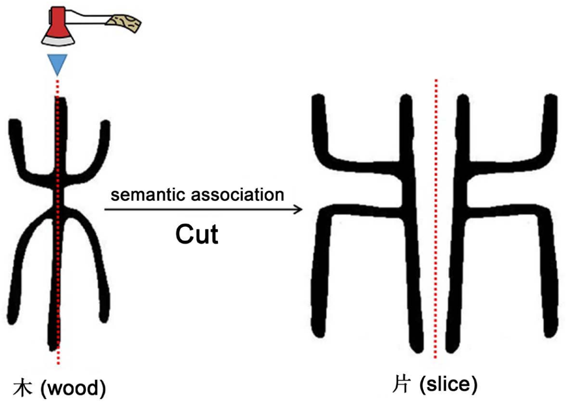


Figure 2. The semantic relation between wood and slice.

semantic knowledge graph (CCSKG), the main task is to discover the entities and the semantic relations between them. HowNet [6] is an online common-sense knowledge base unveiling inter-conceptual relations and inter-attribute relations of concepts as connoting in lexicons of the Chinese and their English equivalents. Its set of sememe is established on a meticulous examination of about 6000 Chinese characters. Therefore, it is a powerful tool for constructing the CCSKG.

Chinese characters are ideographic characters, and radicals are ideographic components. Chinese learners must understand the components and radicals. Xiong *et al.* [7] used radicals to express the semantics of Chinese characters and proposed the concepts of Chinese character genes and families. However, they did not take into account the problem that the same character has different semantics in different contexts.

With the rapid development of deep learning, many research and applications have achieved breakthrough results, such as Oracle Bone Inscription detection [8] [9], corn variety identification [10]. The knowledge base still has its advantages in the era of deep learning. Niu *et al.* [11] verified that integrating word sememe information can improve word representation learning. Xie *et al.* [12] studied the automatic prediction of lexical sememes based on semantic meanings of words encoded by word embeddings. Zeng *et al.* [13] used the sememe information with the attention mechanism to capture the exact meanings of a word, so as to expand and improve the lexicon. However, these studies are advanced applications of Chinese and do not provide a good solution to the problems faced by beginners, especially foreigners, in learning Chinese.

In summary, existing research has not focused on the actual situation of overseas Chinese learners and has not analyzed the differences in the causes of their typos compared to native Chinese speakers. These studies also do not represent Chinese character knowledge from the perspective of their semantics, so they cannot solve the problem of typos caused by similar shapes or pronunciations. This paper constructs a knowledge graph from the perspective of Chinese character semantics and attempts to solve these problems.

The main contributions of this paper are listed as follows:

- Consider the intellectual associations of Chinese characters in terms of their original meanings rather than their glyphs. This makes it easier to solve the difficulties of foreigners in learning Chinese.
- The CCSKG is constructed from the semantic perspective, which makes the entity relations in the graph have rich semantic information. The prototype of the CCSKG comes from the commonly used Chinese characters, which are the basis for foreigners to learn Chinese.
- The scale extension of the CCSKG based on HowNet, which is still built on the basis of semantics. Because HowNet itself is a powerful Chinese-English common sense knowledge base, it has an inherent semantic computing advantage.
- The Chinese sentences are represented as subgraphs in the CCSKG, and typos are corrected by using graph algorithms.
- The elementary level is based on the understanding of Chinese characters rather than words, which greatly reduces the difficulty of learning Chinese.
- The basic requirements of Chinese language learning can be met without large-scale labeled samples and training sets.

The rest of the paper is structured as follows: Section 2 introduces the process of the CCSKG building in detail. Section 3 is the experiment and analysis. The corpus and dataset used are also presented. Finally, Section 4 presents our conclusions and points out the next research priorities.

2. CCSKG Construction

We have presented the construction process of CCSKG in [5]. Firstly, we build an initial set of Chinese characters, called seed set. Secondly, we establish semantic relations between Chinese characters based on the seed set to form the prototype knowledge graph. They are classification clusters based on the Chinese character semantic families. Thirdly, use word pairings in HowNet to expand the radical knowledge graph. Fourthly, based on the similarity calculation of OpenHowNet, more Chinese character entities and relations can be obtained, thereby enriching the knowledge graph. Finally, the integrated entities and relations form the CCSKG. The construction process is shown in **Figure 3**.

2.1. Knowledge Graph Seed Set

There are about 100,000 Chinese characters. Considering the learning and usage scope of overseas Chinese learners, we have collected and organized 6374 commonly used Chinese characters from authoritative Chinese textbooks, including 256 single-component characters. We refer to the set of these Chinese characters as a seed set, where each element is a node in the CCSKG.

2.2. Knowledge Graph Prototype Construction

Sorting the seed set will form a prototype of the knowledge graph. The Chinese characters in the seed set are divided into 190 groups according to 190 radicals, as shown in **Figure 4**. The first Chinese character in each line is a radical, which

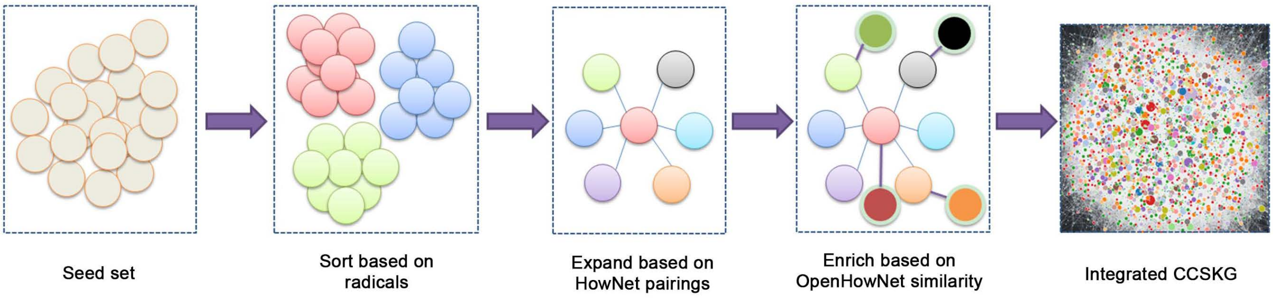
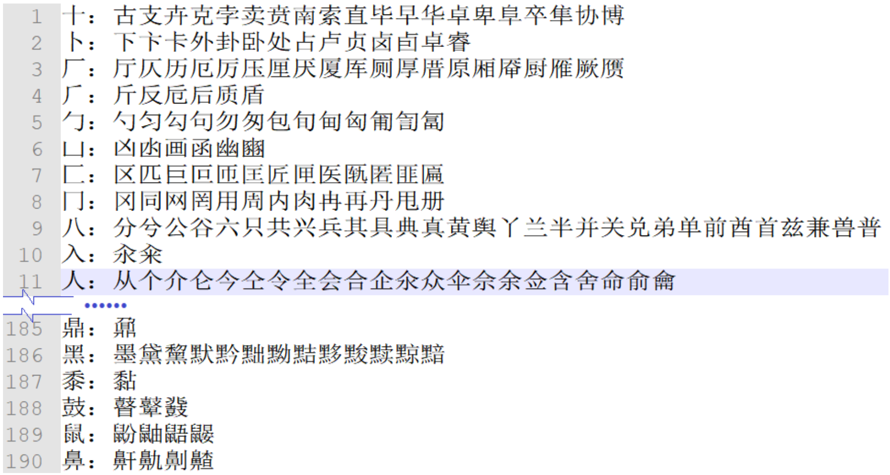

Seed set
Sort based on radicals
Expand based on HowNet pairings
Enrich based on OpenHowNet similarity
Integrated CCSKG

Figure 3. The CCSKG construction process.



```
  1 十： 古支卉克孛卖贲南索直毕早华卓卑阜卒隼协博
  2 卜： 下卞卡外卦卧处占卢贞卤卣卓睿
  3 厂： 厅仄历厄厉压厘厌厦厍厕厚厝原厢厣厨雁厥赝
  4 ⺁： 斤反卮后质盾
  5 勹： 勺匀勾句勿匆包旬甸匈匍訇匐
  6 凵： 凶凼画函幽豳
  7 匚： 区匹巨叵匝匡匠匣医瓯匿匪匾
  8 冂： 冈同网罔用周内肉冉再丹甩册
  9 八： 分兮公谷六只共兴兵其具典真黄舆丫兰半并关兑弟单前酋首兹兼兽普
 10 入： 汆籴
 11 人： 从个介仑今仝令全会合企氽众伞佘余佥含舍命俞龠
      ••••••
185 鼎： 鼐
186 黑： 墨黛黧默黔黜黝黠黟黢黩黥黯
187 黍： 黏
188 鼓： 瞽鼙鼗
189 鼠： 鼢鼬鼯鼹
190 鼻： 鼾鼽劓齄
```

Figure 4. The Chinese characters are divided into 190 groups based on their radicals.

is the basic structural unit of Chinese characters. The characters after the colon in each line are composed of that radical as a component, each line forms a group. Each Chinese character in the same group is related to its radical of the group.

Radicals contain semantic information, and semantic families of Chinese characters can be constructed based on radicals. In this way, we have constructed 190 semantic families. Therefore, the prototype of a knowledge graph of Chinese characters is obtained. The prototype knowledge graph nodes and their corresponding Chinese characters are shown in **Figure 5**.

We notation the original seed character set C. For each Chinese character $c_i \in C, i = 1,2,3,\cdots$, according to the Chinese character radical set $R \in C$, the initial classification is made to form k subsets $F_j \in C, j = 1,2,3,\cdots,k$ named Chinese character semantic families. The reason for the seed set selection is that those commonly used Chinese characters are closest to the daily life, and they are more frequently used and easier to master. There are two reasons for the preliminary classification based on the radicals. The first one is that the radicals and components contain semantic meaning of Chinese characters and the second one is that they are also the basis of current Chinese character learning methods based on glyph. **Algorithm 1** for constructing the semantic families of Chinese characters is as follows.

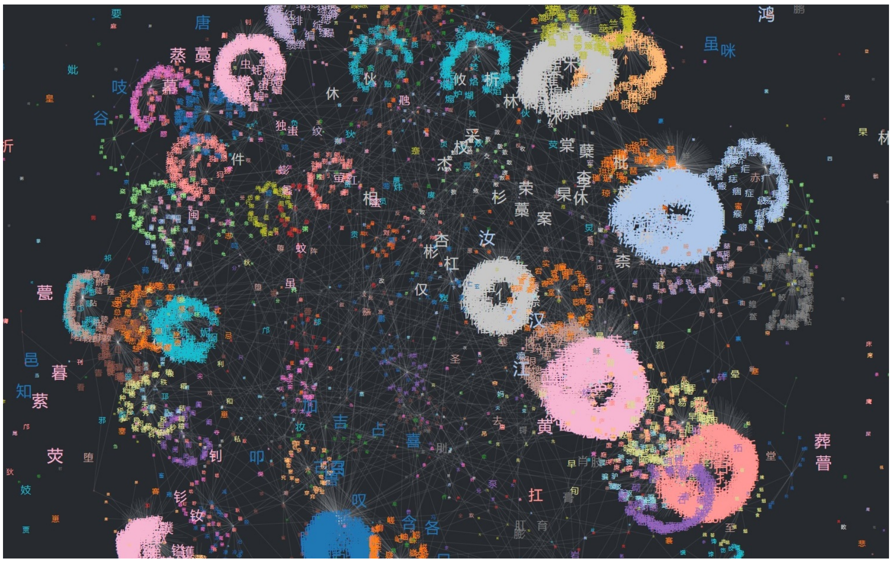


Figure 5. The prototype knowledge graph of Chinese characters.

Algorithm 1 Building the semantic families of Chinese characters.

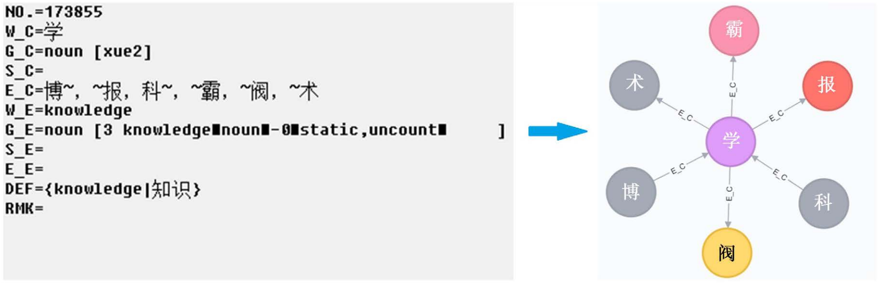
```
Input: The set of Chinese characters, $C_n$; The set of radicals, $R_m$
Output: The set of Chinese character semantic families, $F_m$
$F_m = \Phi$
for each $c_i \in C_n$ do
    $r \leftarrow$ extracting the radical of $c_i$
    if $R_m$ contains $r$ then
        $j \leftarrow$ index of r in $R_m$
        $F_j$ add $c_i$
        $F_m = F_m \cup F_j$
    else
        $R_m$ add $r$
    end if
end for
return $F_m$
```

2.3. Knowledge Graph Extension Based on HowNet Pairings

HowNet provides examples of words that have multiple meanings. These examples emphasize their ability to distinguish rather than their ability to interpret. Their purpose is to provide reliable help for disambiguation. So we can find the correlation between Chinese characters from these examples. Figure 6 shows an example.

In Figure 6, the Chinese character "学(study)" in HowNet provides several examples, by using these examples we can find new Chinese characters as well as the relations between them, thus obtaining a new graph.

The method to extend the knowledge graph based on HowNet is as follows: for each Chinese character $c_{ij} \in F_i$, to find the pairing characters through the word examples in HowNet. These paired characters form a new set W_i. For each $w_{ik} \in W_i$, if $w_{ik} \in C$, the correlation between c_{ij} and w_{ik} is established.



```
NO.=173855
W_C=学
G_C=noun [xue2]
S_C=
E_C=博~，~报，科~，~霸，~阀，~术
W_E=knowledge
G_E=noun [3 knowledge■noun■-0■static,uncount■      ]
S_E=
E_E=
DEF={knowledge|知识}
RMK=
```

(a) Senses of '学' in HowNet

(b) Extension of '学' based on HowNet

Figure 6. An example of extension using HowNet pairings.

If $W_{ik} \notin C$, the W_i is added to the set *C*, then build the relation between c_{ij} and w_{ik}. By using this method, the entities in the knowledge graph of Chinese characters can be expanded, and the more relations between entities can be obtained, so we can get richer semantic information. Since we are concerned about the relations between Chinese characters, we only choose 2-character words as the target when looking for pairing words.

Take any Chinese character node from the prototype of the CCSKG and input it into HowNet to find its senses. By selecting the 2-character words corresponding to the key E_C from the description structure of the senses (see **Figure 6(a)**), more nodes and relations can be obtained based on HowNet's pairings, thereby expanding CCSKG.

2.4. Knowledge Graph Enrich Based on OpenHowNet Similarity

Since most existing machine learning datasets merely provide logical labels, label distributions are unavailable in many real-world applications. Research on label distribution learning (LDL) has gradually attracted attention [14]. Word embedding transforms words into a distributed representation. Therefore, the similarity between words can be obtained through word vectors. Niu *et al.* [11] found that integrating sememe information of Hownet into word representation learning can effectively improve the performance of word embedding. OpenHowNet [15] provides a convenient way to search information in HowNet, display sememe trees, calculate word similarity via sememes, etc. Inspired by this, we also consider fusing distributed representation and knowledge representation to calculate the character semantic similarity.

We extended our calculation model for the correlation of Chinese characters, which is composed of multiple factors. As long as the result of the correlation between two Chinese characters is greater than the specified threshold, they can establish a semantic relation, thereby expanding the CCSKG. The formulas for calculating the relevance of two Chinese characters are as follows.

$$R\left(c_i,c_j\right)=\alpha\cdot\beta\cdot R_{hownet}\left(c_i,c_j\right)+\gamma\cdot Sim_{openhownet}\left(c_i,c_j\right) \tag{1}$$

$$\alpha=\begin{cases}1, & \text{if } w_{ij}\leftarrow c_ic_j \text{ is annotated in HowNet,}\\ 0, & \text{otherwise.}\end{cases} \tag{2}$$

$$\beta = \frac{P_i \cap P_j}{P_i \cup P_j} \tag{3}$$

$$\gamma = \frac{\sum path}{2N_{family}}. \tag{4}$$

where $R(c_i, c_j)$ indicates the semantic correlation degree between the two Chinese characters c_i and c_j; α is called the matching coefficient, which refers to whether c_i and c_j can be combined into a meaningful word. The combination of c_i and c_j is disordered; β is called the component coefficient, which indicates how many common parts of the glyph components that make up c_i and c_j; P_i is the set of components that make up the Chinese character c_i, and P_j is the set of components that make up the Chinese character c_j; γ is referred to as the intimacy coefficient, which represents the semantic distance relationship between the Chinese character and the family of the Chinese character; $\sum path$ represents the sum of the shortest paths between the nearest common ancestors of c_i and c_j; N_{family} means the number of semantic families spanned by $\sum path$.

In formula (1), $R_{hownet}(c_i, c_j)$ represents the semantic relevance of c_i and c_j, which is provided by the HowNet calculation tool. HowNet provides interfaces for semantic relevancy calculations, using these interfaces to compute semantic relevancy between words. If the result is 1, it indicates that there is a correlation between words, so a relation can be established between them. If the result is 0, there is no semantic correlation between them and there is no need to establish the relation. $Sim_{openhownet}(c_i, c_j)$ represents the similarity between c_i and c_j, provided by OpenHowNet, its algorithm implementation is based on [16].

2.5. CCSKG Integration

The integration of CCSKG is to integrate the relations based on partial classification and the new entities and relations based on HowNet and OpenHowNet extension, and remove duplicates in both methods. All Chinese characters in character set C are regarded as nodes, and the relations between characters as edges, thus forming the semantic knowledge graph of Chinese characters. The knowledge graph based on semantic correlation can intuitively display the related Chinese characters. Taking these related Chinese characters as the knowledge community, they can be grasped and understood semantically, thus effectively avoiding the problem of learning Chinese characters based on glyph discrimination.

The expansion of pairing words and semantic relevancy calculation based on HowNet and OpenHowNet greatly enriches the prototype of CCSKG. Comparing the prototype knowledge graph with the integrated one, we found that the nodes and relations in the knowledge graph have increased dramatically. The number of nodes increased by 5.93% and the number of relationships increased by 400%. As shown in **Figure 7**.

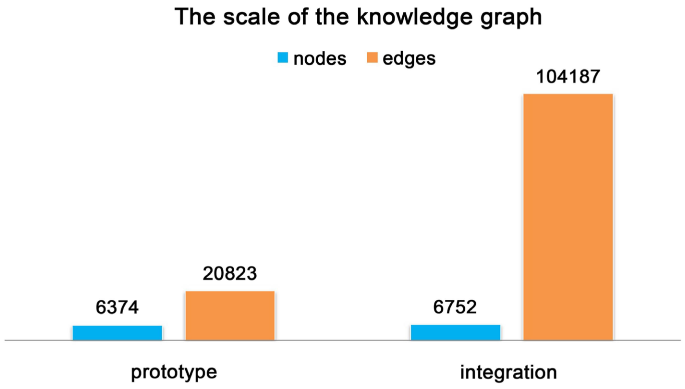


Figure 7. The scale comparison before and after integration.

The CCSKG consists of Chinese characters as nodes, and the relations between the nodes are directional, which indicates the collocation of the characters with their radicals and with other Chinese characters. To facilitate queries and semantic computations, we store the CCSKG in the Neo4j graph database. The current size of the CCSKG is 6752 nodes and 104,187 relations. A screenshot of the CCSKG fragment is shown in Figure 8.

2.6. Link Prediction Algorithm for Chinese Correction

Once the CCSKG is constructed, we can perform Chinese correction based on the graph structure. It is regarded as a link prediction problem. That is, the set of Chinese characters that constitute meaningful words in a sentence is regarded as a subgraph in the CCSKG, and the correctness of the target word is determined by computing whether the characters that make it up belong to the same community as that subgraph. Since link prediction algorithms help determine the closeness of a pair of nodes using the topology of the graph. The computed scores can then be used to predict new relationships between them [17].

Specifically, given a CCSKG, denoted by $G=(E,R,F)$, where E denotes the set of nodes, R denotes the set of relations, and F denotes the set of facts. A Chinese sentence can be converted into a subgraph $G_{sub}=(E_c,R_c,F_c)$ of G, where E_c is the set of Chinese characters, R_c is the set of relations among Chinese characters, and F_c is the set of facts composed of Chinese characters. The G_{sub} composed by a sentence can be regarded as a community, and our approach is to calculate whether the candidate Chinese characters belong to the same community to determine whether the word they constitute is a correct word. For the sake of simplicity, we assume that there is only one word in a Chinese sentence that needs to be judged as correct or not. The calculation process is as described in **Algorithm 2**. Thanks to Neo4j for providing Graph Data Science (GDS) library that can provide us with Community detection and similarity algorithms.

3. Experiments and Analysis

3.1. Dataset

The dataset used in this paper comes from the HSK dynamic composition corpus

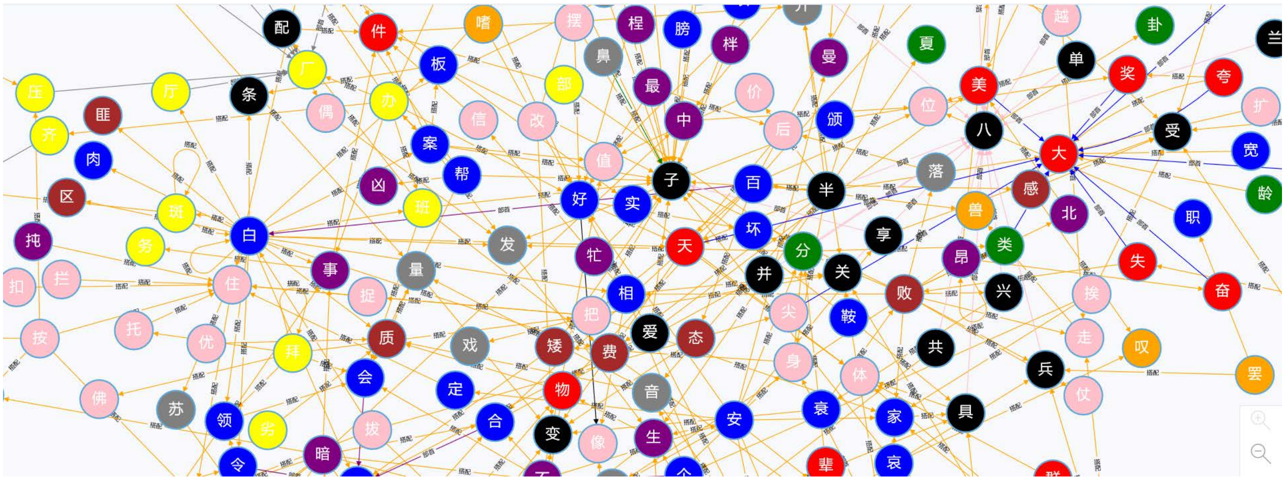


Figure 8. The screenshot of the CCSKG.

Algorithm 2 Chinese correction based on CCSKG.

Input: The knowledge graph $G = (E, R, F)$; the Chinese sentence S; the predicted 2-character word $w = c_1c_2$; Neo4j Graph Data Science Library *gds.alpha*.

Output: The corrected Chinese characters $w' = c'_1c'_2$

```
$G_{sub} = \Phi$
$S.tokenizer()$
$W \leftarrow S.removeStopwords()$
for each $w_i \in W$ do
    $C \leftarrow w_i.toCharArray()$
    for each $z_i \in C$ do
        if $z_i$ matched in $G$ then
            $G_{sub}$ add $z_i$
        end if
    end for
end for
if $c_1 \in E$ and $c_2 \in E$ and $r(c_1, c_2) \in R$ then
    $score \leftarrow alpha.linkprediction.sameCommunity(c_1, c_2)$
    if $score = 1.0$ then
        $G_{sub}$ add $c_1$
        $G_{sub}$ add $c_2$
        $c'_1c'_2 \leftarrow c_1c_2$
    end if
else
    $\arg\max_{t_1,t_2 \in E} f(t_1, t_2) = \{t_1, t_2 \in E : f(t_1, t_2) = alpha.linkprediction.totalNeighbors(t_1, t_2)\}$
    $c'_1c'_2 \leftarrow t_1t_2$
end if
return $c'_1c'_2$
```

version 1.1 created by Beijing Language and Culture University, which is a corpus of compositions written by foreigners whose native language is not Chinese for the Advanced Chinese Proficiency Test. The compositions of some foreign students from 1992 to 2005 are collected. The scale of the corpus is 11,569 articles and 4.24 million characters in 29 composition topics [18]. The corpus provides two versions: annotated corpus and original corpus. The annotated one is a corpus that is manually entered into a computer and manually marked with various interlanguage errors. The original one refers to the electronically scanned corpus of the students' original compositions.

The dataset was generated based on the following considerations: 1) since we mainly focus on the correction of mispronounced characters, we chose the annotated mispronounced characters corpus from the HSK dynamic composition corpus; 2) we have separated the correct Chinese characters and mispronounced characters, since the original annotated sentences in the corpus merges the two together.

There are several dataset samples shown in **Table 3**.

In **Table 3**, the *annotated sentences* are the original sentences in the HSK composition corpus, and the characters marked with B indicate that they are mispronounced characters; the *source sentences* represent the actual composition sentences of the foreign students, which may contain errors or may be correct; the *target sentences* are correct, they may be the same as the *source sentences*, or they may be manually corrected sentences. It is worth noting that we are concerned about the B-marked characters in the HSK corpus, and other marked characters need to be processed accordingly. For example, the Chinese character "慮" marked F is a traditional Chinese character, it is semantically correct, but the glyph is different from the simplified Chinese character "虑". Therefore the target sentences will contain both glyphs. See sentences No. 2 and No. 3 in **Table 3**.

3.2. Chinese Character Typos Correction

To verify the validity of the CCSKG, we also conducted experiments with word correction as the task. Among the experimental data, the probability of incorrect sentences is 14.67%. The experimental task is described as giving the target word that needs to be detected, finding its corresponding position in the experimental sentences, removing it through MASK, and transforming it into a cloze task. The predicted result is considered as the corrected word and then compares with the word dropped by MASK. Since our current CCSKG considers 2-character words, our word correction targets are also 2-character words. We employ False Acceptance Rate (FAR) and Recall as evaluation criteria. The formulas are as follows:

$$\text{FAR} = \frac{N_{wc}}{N_c} \tag{5}$$

$$\text{Recall} = \frac{N_{er}}{N_e} \tag{6}$$

where N_{wc} is the number of sentences that were correct but were incorrectly corrected; N_c is the number of correct sentences originally; N_{er} means the number of sentences that were originally incorrect and were finally corrected correctly; N_e is the number of incorrect sentences originally.

We compare the effect of our proposed method with keras-bert and pycorrector-ernie. BERT [19] is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all

Table 3. Dataset samples.

No.	annotated sentences	source sentences	target sentences
1	我觉得如果要打工，那么应该考虑周围的环[B 坏]境。	我觉得如果要打工，那么应该考虑周围的坏境。	我觉得如果要打工，那么应该考虑周围的环境。
2	物质都足够了以后人还要考虑[F 慮]这一点。	物质都足够了以后人还要考慮这一点。	物质都足够了以后人还要考虑这一点。
3	物质都足够了以后人还要考虑[F 慮]这一点。	物质都足够了以后人还要考慮这一点。	物质都足够了以后人还要考慮这一点。
4	你为什么不考虑我的意见呢？	你为什么不考虑我的意见呢？	你为什么不考虑我的意见呢？

layers. ERNIE [20] is designed to learn language representation enhanced by knowledge masking strategies, which includes entity-level masking and phrase-level masking. Pycorrector [21] is a Chinese text error correction tool. It uses the language model to detect errors, Chinese Pinyin feature and shape feature to correct Chinese text error, it can be used for Chinese Pinyin and stroke input method. The experimental results are shown in **Table 4**.

It can be seen from **Table 4** that our method has achieved the best results. It has the lowest FAR and the highest Recall. We are more concerned with the whole word mask, compared to the best results of keras-bert and pycorrector + ernie, our FAR is reduced by 38.28%, and the Recall is increased by 40.91%. The analysis of the experimental results is as follows:

- Incorporating knowledge into self-supervised learning methods can indeed improve the performance of natural language processing tasks. For example, compared to keras-bert and pycorrector + ernie our method incorporates the knowledge elements of HowNet.
- The importance of the 2 characters that make up the word is significantly different, and predicting the latter character by the former word is significantly better than predicting the former character by the latter word. This also proves that the relationship between words in the CCSKG is directional.
- Since the HSK composition corpus comes from the compositions of international students, there are some grammatical errors in the sentences, which affect the performance of typos correction.
- The accuracy of word segmentation tools has an impact on text error correction. For example, “世界/上” is mistakenly divided into “界上”, “现在/考虑” is mistakenly divided into “在考” and so on.
- Bert and Ernie provide large-scale pre-training corpora, but these corpora are from high-quality Chinese material, which is significantly different from the HSK composition corpus. In the absence of large-scale HSK pre-training corpora, incorporating knowledge as a guide is an effective solution.
- HSK composition corpus contains traditional Chinese characters, such as “考慮”. In experiments, we found that keras-bert and our method have poor performance in processing traditional characters. In the case of pycorrector, although it supports traditional Chinese characters, it does not always convert them to the correct simplified Chinese and often substitutes them based on Pinyin, which leads to errors.

Table 4. Typos correction results.

	first character mask		second character mask		the whole word mask	
	FAR (%)	Recall (%)	FAR (%)	Recall (%)	FAR (%)	Recall (%)
keras-bert	12.50	22.73	1.56	95.45	83.59	4.55
pycorrector + ernie	5.47	4.55	6.25	95.45	86.72	13.64
ours	3.91	59.09	2.34	95.45	45.31	54.55

- When correcting typos, the methods used in the experiment have a lower accuracy in correcting homophones than those caused by similar shapes. Adding pronunciation attributes to CCSKG may be a good solution.

4. Conclusions

To help foreign learners to catch the meaning of Chinese characters during the Chinese foreign communication and propagate, the CCSKG construction method is proposed. Unlike other Chinese character glyph description methods, we pay close attention to the Chinese characters' original meaning and their semantic relevancy. In addition, as a powerful knowledge base, HowNet is used to extend and perfect the knowledge graph. We have obtained 6752 Chinese characters and 104,187 relations to meet the needs of Chinese overseas communication and international education. We realized the visualization of the CCSKG and verified its effectiveness in word correction through experiments. However, HowNet currently does not collect traditional Chinese characters and words, so it is temporarily unable to perform semantic analysis on traditional Chinese characters which are often encountered in overseas Chinese communication scenes. Moreover, currently, our CCSKG only considers 2-character words. In future work, we will integrate other resources of Chinese character knowledge into semantic representation models, such as Oracle Bone Inscriptions characters, traditional Chinese characters. And extend them to the CCSKG. And the representation of multi-character words is also an issue that needs to be studied. Thus, we can better serve Chinese cultural exchange and overseas dissemination.

The earlier simple version of this paper was presented at the 2022 International Conference on Computer Engineering and Artificial Intelligence.

Acknowledgement

This work is supported by the Key Technology Project of Henan Educational Department of China (22ZX010), the National Natural Science Foundation of China (62106007), the Chinese Ministry of Education and National Language Commission Special Project for Research and Application of Oracle Bone Inscriptions and other ancient characters (YWZ-J023), the Industry-University-Research Innovation Fund of China (2021RYA05002), and Henan Provincial Department of Science and Technology Research Project (232102320169).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Wang H.-B. (2014) Mexican Chinese Learners Cognitive Process of Alphabetic and Ideographic Writings. *Journal of Ningbo University* (*Educational Science Edition*), **4**, 1-6.

[2] Chou, Y. M. and Huang, C.-R. (2013) The Formal Representation for Chinese Characters. *Contemporary Linguistics*, **15**, 142-161.

[3] Yaowen Jiaozi Editorial Office (2006) The 100 Most Common Typos in Contemporary Chinese Publications. *Yaowen jiaozi*, **6**, 4-6.

[4] Zhang, B.L. (2010) Avoidance and Overgeneralization-An Investigation of Acquisition of the Ba-Sentence Based on the HSK Dynamic Composition Corpus. *Chinese Teaching in the World*, 2.

[5] Xiong, J., Liu, G., Guo, T. and Chen, Y. (2022) Construction of Chinese Character Semantic Knowledge Graph for Overseas Chinese Learners. *Proceedings of* 2022 *International Conference on Computer Engineering and Artificial Intelligence* (ICCEAI), Shijiazhuang, 22-24 July 2022, 153-157.
https://doi.org/10.1109/ICCEAI55464.2022.00040

[6] Dong, Z., Qiang, D. and Hao, C. (2010) HowNet and Its Computation of Meaning. *COLING* 2010, 23*rd International Conference on Computational Linguistics, Demonstrations Volume*, 23-27 August 2010, Beijing. Association for Computational Linguistics.

[7] Xiong, J., Liu, X. and Li, Q. (2015) Ontology Description of Chinese Character semantics. *Proceedings of* 2015 *IEEE International Conference on Computer and Information Technology*, *Ubiquitous Computing and Communications*, *Dependable*, *Autonomic and Secure Computing*, *Pervasive Intelligence and Computing*, Liverpool, 26-28 October 2015, 709-713.
https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.103

[8] Xing, J., Liu, G. and Xiong, J. (2019) Oracle Bone Inscription Detection: A Survey of Oracle Bone Inscription Detection Based on Deep Learning Algorithm. *Proceedings of the International Conference on Artificial Intelligence*, *Information Processing and Cloud Computing* (*AIIPCC* 19), New York, 19-21 December 2019, 1-8.
https://doi.org/10.1145/3371425.3371434

[9] Xiong, J., Liu, G., Liu, Y. and Liu, M. (2021) Oracle Bone Inscriptions Information Processing Based on Multi-Modal Knowledge Graph. *Computers & Electrical Engineering*, **92**, Article ID: 107173.
https://doi.org/10.1016/j.compeleceng.2021.107173

[10] Zhang, W., Li, Z., Sun, H., Zhang, Q., Zhuang, P. and Li, C. (2022) SSTNet: Spatial, Spectral and Texture Aware Attention Network Using Hyperspectral Image for Corn Variety Identification. *IEEE Geoscience and Remote Sensing Letters*, **19**, 1-5.
https://doi.org/10.1109/LGRS.2022.3225215

[11] Niu, Y., Xie, R., Liu, Z. and Sun, M. (2017) Improved Word Representation Learning with Sememes. *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Vancouver, 30 July-4 August 2017, 2049-2058. https://doi.org/10.18653/v1/P17-1187

[12] Xie, R., Yuan, X., Liu, Z. and Sun, M. (2017) Lexical Sememe Prediction via Word

Embeddings and Matrix factorization. *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Melbourne, 19-25 August 2017, 4200-4206. https://doi.org/10.24963/ijcai.2017/587

[13] Zeng, X., Yang, C., Tu, C., Liu, Z. and Sun, M. (2018) Chinese LIWC Lexicon Expansion via Hierarchical Classification of Word Embeddings with Sememe Attention. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, Online, 22 February-1 March 2022, 5650-5657. https://doi.org/10.1609/aaai.v32i1.11982

[14] Zheng, Q., Zhu, J., Tang, H., Liu, X., Li, Z. and Lu, H. (2021) Generalized Label Enhancement with Sample Correlations. *IEEE Transactions on Knowledge and Data Engineering*, **35**, 482-495. https://doi.org/10.1109/TKDE.2021.3073157

[15] Qi, F., Yang, C., Liu, Z., Dong, Q., Sun, M. and Dong, Z. (2019) Openhownet: An Open Sememe-Based Lexical Knowledge Base. (Preprint)

[16] Liu, J., Xu, J. and Zhang, Y. (2013) An Approach of Hybrid Hierarchical Structure for Word Similarity Computing by Hownet. *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, 14-19 October 2013, 927-931.

[17] Needham, M. and Hodler, A.E. (2019) Graph Algorithms: Practical Examples in Apache Spark and Neo4j. O'Reilly Media, Sebastopol.

[18] Cheng, S.-M., Yu, C.-H., and Chen, H. (2014) Chinese Word Ordering Errors Detection and Correction for Non-Native Chinese Language Learners. *Proceedings of COLING* 2014 25*th International Conference on Computational Linguistics*: *Technical Papers*, Dublin, 23-29 August 2014, 279-289.

[19] Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Vol. 1 (Long and Short Papers), Minneapolis, 2-7 June 2019, 4171-4186.

[20] Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H. and Wu, H. (2019) Ernie: Enhanced Representation through Knowledge Integration. (Preprint)

[21] Xu, M. (2021) Pycorrector: Text Error Correction Tool. https://github.com/shibing624/pycorrector