ZiNet: Linking Chinese Characters Spanning Three Thousand Years

Yang Chi[1], **Fausto Giunchiglia**[1,2,3], **Daqian Shi**[3], **Xiaolei Diao**[3]
Chuntao Li[4], **Hao Xu**[1,2,*]
[1]School of Artificial Intelligence, Jilin University, Changchun, China
[2]College of Computer Science and Technology, Jilin University, Changchun, China
[3]DISI, University of Trento, Trento, Italy
[4]School of Archaeology, Jilin University, Changchun, China
yangchi19@mails.jlu.edu.cn, {xuhao,lct33}@jlu.edu.cn
{fausto.giunchiglia,daqian.shi,xiaolei.diao}@unitn.it

Abstract

Modern Chinese characters evolved from 3,000 years ago. Up to now, tens of thousands of glyphs of ancient characters have been discovered, which must be deciphered by experts to interpret unearthed documents. Experts usually need to compare each ancient character to be examined with similar known ones in whole historical periods. However, it is inevitably limited by human memory and experience, which often cost a lot of time but associations are limited to a small scope. To help researchers discover glyph similar characters, this paper introduces ZiNet, the first diachronic knowledge base describing relationships and evolution of Chinese characters and words. In addition, powered by the knowledge of radical systems in ZiNet, this paper introduces glyph similarity measurement between ancient Chinese characters, which could capture similar glyph pairs that are potentially related in origins or semantics. Results show strong positive correlations between scores from the method and from human experts. Finally, qualitative analysis and implicit future applications are presented.

1 Introduction

The evaluation of Chinese character can be divided into two stages: the ancient stage (before Han dynasty, 202 BC) and the clerical and standard script stage (after Han dynasty) (Qiu et al., 2000). At the former stage, ancient characters do not have a fixed shape, and their glyphs show several differences respect to modern characters: representatives include the Oracle bone script (Oracle) in the Shang Dynasty (about 1300 BC), which appears on animal bones or turtle shells (Boltz, 1986), the Chinese bronze script (Bronze, about 1000 BC) appeared on bronze wares (Shaughnessy, 1991) and the script belonging to the Warring States period (States), mainly recorded on wooden slips (about 400 BC) (Qiu, 2014). Evolution of glyphs of Chinese characters can be observed in Figure 1.

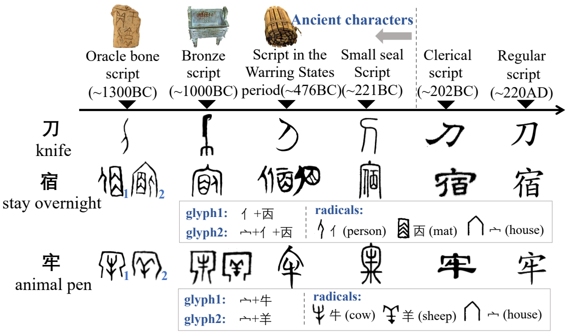

Figure 1: Examples of historical evolution of Chinese characters associated with unfixed glyphs and radical compositions (the pictures on the top show unearthed ancient characters respectively written on the turtle shell, bronze ware and wooden slips).

Ancient unearthed documents show a wealth of information regarding that historical periods (Boltz, 1986; Shaughnessy, 1991; Qiu, 2014), which is great significant for understanding the culture and history of China, as well as the whole world. Nevertheless, nearly half of ancient characters cannot be deciphered yet. The purpose of deciphering ancient character is to find the modern Chinese characters evolved from it and give enough interpretations and evidences in terms of glyphs, phonetics and semantics. According to the systematic nature and evolution law of Chinese characters, experts need to compare the character to be examined with similar known characters in history. However, there are tens of thousands of glyphs of characters appeared in history, discovering similar character heavily relies on expert experience, which is inevitably limited by human memory and reduces the comprehensiveness and efficiency.

To measure similarity between ancient characters, automatic methods face challenges: (1) it lacks available resources of ancient Chinese, which means existing algorithms, especially supervised al-

[*] Corresponding author

Findings of the Association for Computational Linguistics: ACL 2022, pages 3061 - 3070
May 22-27, 2022 ©2022 Association for Computational Linguistics

gorithms cannot be directly used to solve this problem. And ancient characters do not have features such as standard code, pinyin and strokes, which have been widely used for describing the modern one. (2) it is complicated to represent and measure ancient characters. For instance, edit distance is widely used to measure orthographic similarity between words in Romance languages; however, it is not suitable for measuring glyph similarity between pictographic Chinese characters.

Based on above considerations, the main contributions of this paper are: (1) introduces ZiNet, the first diachronic knowledge base for linking Chinese characters and words across various historical periods. (2) as the first application of ZiNet, this paper introduces methods for glyph similarity measurement, which aims at giving glyph similar scores for ancient Chinese character pairs.

There are two main characteristics of ZiNet comparing to existing lexical resources: (1) it is designed based on the systematic nature of Chinese characters. The smallest unit of ZiNet is the radical, the component of character, which is of significance for analyzing semantics or phonics of characters (details will be discussed in Section 3.1). (2) ZiNet is diachronic, which integrates characters and words across historical periods, and aims to portray their evolution. Powered by knowledge of ZiNet, our glyph similarity measurement method could capture the glyphs that are potentially relevant in terms of origins or semantics, which is meaningful in researches of Chinese characters. Results shown a strong positive correlation between the methods scores and human experts.

The paper is organized as follows: Section 2 presents the state of the art; Section 3 describes key information of ZiNet; Section 4 describes glyph similarity measurement; results are proposed in Section 5; Section 6 shows implications and future works; Section 7 concludes the paper and Section 8 shows ethics.

2 State of the Art

2.1 Processing Ancient Chinese Character

Zhang et al. (2020) built a real-world dataset OB-Rejoin, which proposes an effective algorithm to rejoin Oracle fragments. Han et al. (2020) proposed an Oracle information system, known as IsOBS, which records Oracle rubbings, documents, Oracle characters and all their variants. Jiao et al. (2021) generated a network for Oracle characters according to their structures and documents. They classified the semantically-similar Oracle characters by analyzing the network module.

2.2 Lexical Resources and Cognate Discovery

WordNet-oriented (George, 1995) lexical resources are widely used in NLP tasks. Their architecture consists in synset as basic semantic units to integrate words senses, which are related to each other, thus forming a conceptual semantic network. Multilingual resources, such as Open Multilingual WordNet (Bond and Foster, 2013), BabelNet (Navigli and Ponzetto, 2012) and Universal Knowledge Core (Giunchiglia et al., 2017), which integrated words and concepts from all over the world, can support NLP tasks in languages that lack resources.

According to historical linguistics, cognate identification needs to consider three dimensions: semantic, phonetic and orthographic similarity (Arnaud et al., 2017), who dealt with researching ancient Chinese. Hauer and Kondrak (2011) designed rich set of features to capture similarity. Batsuren et al. (2020) considered evidence in the form of a combined orthographic and geographic relatedness. Snyder et al. (2010) designed a Bayesian model to incorporate linguistic constraints, which includes customized priors for alphabet matching and morphological structure. Luo et al. (2019) automatically deciphered ancient languages by evaluating the accuracy of aligning words from a lost language to their counterparts in a known language. According to these works, orthographic similarity is an important indicator; however, measurements like edit distance cannot be directly applied to Chinese characters.

3 ZiNet

3.1 Motivation

ZiNet has been created in order to link Chinese characters and words in history, according to their glyphs, semantics and phonetics to support knowledge-powered algorithms during processing of Chinese or ancient Chinese. Here we will give a general outline of key knowledge to help understand the structure of ZiNet and the reasons why it has been developed.

Relation between word and character: Chinese words are composed of one or more characters; the latter can also be regarded as monosyllable words when expressing semantics. For example, character (or monosyllabic word) "宿"

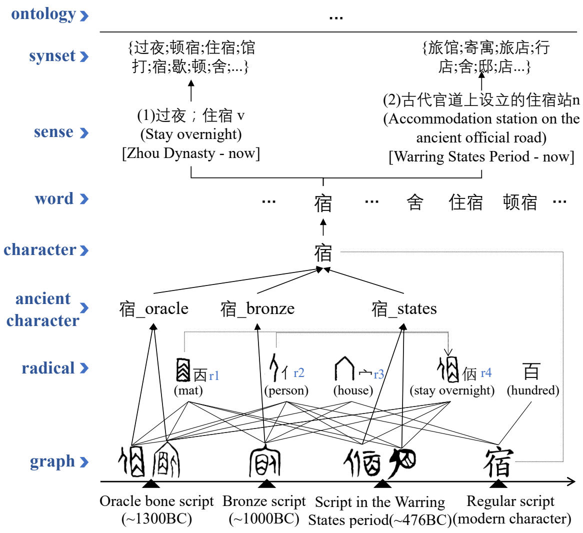

Figure 2: Structure of the ZiNet.

(stay overnight) in Figure 1 can participate in the formation of the polysyllabic word "住宿" (get accommodation).

Relation between character and radical: radical is the conventional structural unit that can participate in the characters composition; radicals themselves are also characters, or variants of characters. For instance, in ZiNet, the radical of the single modern character "刀" (knife) is itself "刀", and the radicals of the compound modern character "宿" (stay overnight) are "宀" (house), " 亻 " (person) and "百" (hundred).

Radical and deciphering: Radicals knowledge is crucial for related researches, because radical is related to the phonetics or semantics of the character. For instance, "宀" (house), " 亻 " (person) are related to the semantics of "宿" (stay overnight). Thus, through radicals and relationships between them, experts are able to discover further phonetic or semantic related characters that may implicit clues for deciphering.

Evolution: The glyph of character is evolving through historical periods. For instance, Figure 1 shows the radicals of Oracle character "宿". In that ancient period, the bottom-right radical of "宿" is not "百" (hundred), but another similar character that means "mat". these objects should be represented within a diachronic network, in order to explore their implicit evolution rules.

3.2 Structure of ZiNet

In the current stage, ZiNet is composed by seven layers and there are relations between layers (Figure 2); in the future, an eighth layer of Ontology is aimed to be added, in order to describe human life through varied historical periods, by linking synsets to concepts and topics.

- **Glyph:** Character writing shapes. ZiNet integrates rubbing images from unearthed artifacts for each glyph.

- **Radical:** The components of character. In ZiNet, all glyphs are associated with corresponding radicals at two levels of granularity (Compound radicals can be further split into finer-grained units. For instance, in Figure 2, r_4 is a compound radical, consisting of r_1 and r_2.).

- **Ancient Character:** Chinese characters in ancient historical periods. All ancient glyphs should be associated with the corresponding ancient character.

- **Character:** Including deciphered and undeciphered characters: the former is further divided into modern and dead character. Ancient characters belonging to different periods that represent the same character should be linked. If ancient character is deciphered, and is being currently used, it should be linked to modern character. Else, if the ancient character is deciphered but is not used, it should be linked to the corresponding dead character. Finally, undeciphered ancient characters should be linked to the corresponding undeciphered character.

- **Word:** Mono-syllable (character) and multi-syllable word in Chinese history.

- **Sense:** Meaning of word. All words should be associated with their corresponding senses.

- **Synset:** A set of at least one synonym. All senses should be associated with the corresponding synset.

The organization of Word, Sense, Synset layers are designed based on WordNet. One word might have several senses; senses associated to the same meaning are linked to the same synset.

Other bottom of layers is different with existing lexical resources, which are designed following the systematic nature of Chinese character. In order to research ancient Chinese characters, knowledge on glyphs and radicals must be explicitly provided.

The other key characteristic is diachronism, which is reflected in two ways: (1) At the glyph level, ZiNet aims at covering the critical period of evolution of Chinese characters: Oracle, Bronze, States and modern characters have been integrated up to now. (2) At the sense level, for each sense, the earliest and latest dynasty where it appeared are annotated, according to the records provided by dictionaries.

Here we introduce two relations inside ancient character layer, which are used to measure glyph similarity:

- **Derivation (分化):** A proliferation phenomenon of Chinese characters: based on a certain glyph of a mother character, making one or several new characters that are glyph-consistent and related to the semantics of the mother character. In ZiNet, if character B (e.g., "東" (bag)) is derived from character A (e.g., "束" (tie)), there would be a Derivation relation between them.
$ancient\_char(B) \xrightarrow{D} ancient\_char(A)$.

- **Indication (指事):** An abstract method to create a new Chinese characters by directly adding a indicative symbol on a specific position of the glyph of the mother character, the new character meaning is related to the position indicated by symbol. If a new character B (e.g.,"刃" (knife edge)) is created by adding a symbol on the specific position(e.g., edge) of a pictographic character A(e.g., "刀" (knife)), there would be an Indication relation between them.
$ancient\_char(B) \xrightarrow{I} ancient\_char(A)$.

3.3 Statistics of ZiNet

ZiNet is constantly developing. All characters, glyphs and rubbings images were provided by experts on Chinese characters. Radicals of each ancient character and relations were also been split and proofread by experts, who referred to dozens of authoritative publications, of which the most representatives are (Chinese Academy of Social Science (CASS), 1984) and (Guo and Hu, 1978). Most of the words and senses in ZiNet were acquired from authoritative ancient dictionaries, such as Shuowenjiezi (Shen Xu, 1963) and a few original senses in far ancient periods were provided by experts. Synsets were automatically associated according to the definitions of the senses. Up to now, ZiNet includes three historical Chinese periods: Oracle, Bronze, and States. Table 1 lists statistical information. ZiNet is extensible, as Figure 2 shows, the Glyph and Ancient Character layers are independent for each historical period, which allows it to conveniently extend to other historical periods in future.

Object	**Statistics**
Rubbing image	15175=Oracle; 14289=Bronze; 28421=States
Glyph	2913=Oracle; 3225=Bronze; 7232=States
Radical	584=Oracle; 853=Bronze; 868=States
Ancient character	2543=Oracle; 2319=Bronze; 5632=States
Character	Deciphered character: 1283=Oracle; 2466≤ Bronze; 4478≤States; 18966≤Present Undeciphered character: 1260=Oracle; 1714≤ Bronze; 4118≤ States
Word	423997

Sense	69825≤206BC;
	177570 ≤ 618AD;
	315181 ≤ 1368AD;
	386949≤1840AD;
	570764≤ Present

Synset	366544

Table 1: ZiNet statistics ("=" means an object existed in that historical period; "≤" means an object had appeared before, or during that period).

4 Glyph Similarity Measurement

4.1 Key Points of Glyph Similarity

The task is to give glyph-similar scores for each ancient character pair: this does not only include the pictographic similarity of character shapes, but also between their radical systems. In this paper, we consider the following four points:

(1) Similar character shape: Two pictographic characters have similar shapes. For example, the pictographic character "刀"(knife) in Figure 1 is depicted in the form of a knife. If the shape of another character also resembles a knife, they are defined as glyph similar.

(2) Sharing radicals: Two characters sharing radicals. For instance, the character "宿" (stay overnight) in Figure 2 is formed by radicals ac-

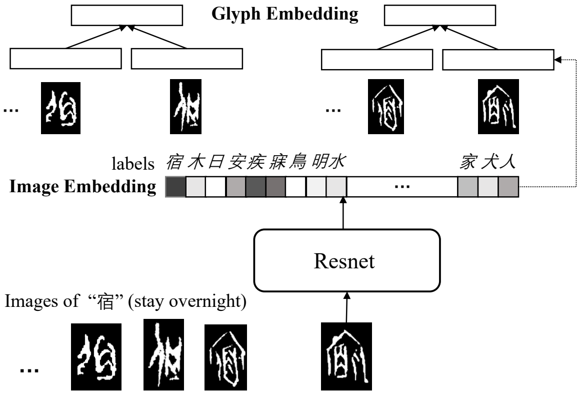
Glyph Embedding
...
...
labels 宿 木 日 安 疾 寐 鳥 明 水
家 犬 人
Image Embedding
...
Resnet
Images of “宿” (stay overnight)
...

Figure 3: Procedure to generate glyph embedding for pictographic similarity.

cording to their respective meanings, rather than by directly drawing the object. If one character shares radicals with it, they are defined as glyph similar.

(3) There are Derivation or Indication relations between their radicals: In general cases, characters do not share radicals; however, their radicals are related in Derivation or Indication (Section 3.2). If two characters respectively contain related radicals, they are defined as glyph similar.

(4) Their radicals are universal when composing a character: In other cases, radicals of two characters do not have relations; however, when composing a character, they are universally used to show the same semantics. Universal radical pairs can be automatically discovered, by exploring radical pairs that are mutually substituted in synchronic, or diachronic different glyphs of the same character in ZiNet. For example, in Figure 1, the character "牢" (animal pen) has two different Oracle glyphs: the first contains the radicals of "宀" (house) and "牛" (cow), whereas the second contains "宀" (house) and "羊" (sheep). In this case, "牛" (cow) and "羊" (sheep) is a pair of synchronic substitutable radicals. We consider the characters respectively containing them as glyph similar.

Pictographic Similarity (PicSim), Radical LCS Similarity (RLCSSim) and Graph Similarity (GraphSim) will be introduced respectively in Section 4.2, 4.3 and 4.4, respectively. While the former aims at measuring the similarity between character shapes, RLCSSim and GraphSim focus on measuring similarities between radical systems.

4.2 Pictographic Similarity

The intuition to measure the similarity of pictographic characters is to consider them as pictures. Deep Residual Network (ResNet) (He et al., 2016) is used to obtain the high-dimensional vector of images of ancient characters, as shown in Figure 3. There are n ancient characters and m images of characters in total; the set of images is $X(x_1, x_2, \dots x_m)$, and that of characters is $C(c_1, c_2, \dots c_n)$. The network task is to classify each image x into the corresponding character c, $p(c|x, \varphi)$ is used to denote the probability that an image x belongs to the character c, where φ is the parameter that needs to be trained to acquire. The network input is the image x, while the output is the $|C|$-dimensional vector: each dimension represents the probability p of each character label c. At the training step, images and their associated Chinese character labels are provided. We minimize cross-entropy loss function to get the optimal parameters φ.

The $|C|$-dimensional vector output is then directly used as the image embedding, $\vec{I}$. As a next step, given the set $ImageSet$ that contains all images belonging to glyph g, the glyph embedding, $\vec{G}$ of g, is set to the average of embedding of images in $ImageSet$.

$$\vec{G_i} = \frac{1}{|ImageSet_i|} \sum_{x_j \in ImageSet_i} \vec{I_j} \tag{1}$$

After obtaining the glyph embedding $\vec{G}$, cosine similarity is used to get the similarity between glyph pairs. It is multiplied by a hyper-parameter α here. Only when two glyphs share the same or related radicals, then $\alpha = 1$, otherwise, α will be set as a value greater than 0, and less than 1. $RSet_i$ consists in the collection of radicals, and their related radicals (derivative, indicative or universal relations as introduced in Section 4.1) of g_i.

$$Sim(g_i, g_j) = \alpha Cosine(g_i, g_j), \tag{2}$$

$$\begin{cases} \alpha = 1, & RSet_i \cap RSet_j \neq \emptyset \\ 0 < \alpha < 1, & Otherwise \end{cases}$$

Finally, given the $GlyphSet$ that contains all glyphs belonging to character c, the PicSim between two characters is the maximum similarity of the combination between their glyph pairs.

$$PicSim(c_k, c_g) = Max\{Sim(g_i, g_j)\}, \tag{3}$$
$$(g_i \in GlyphSet_k, g_j \in GlyphSet_g)$$

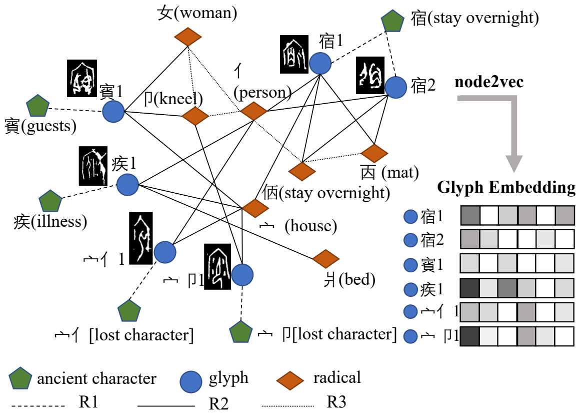

Figure 4: Procedure to generate glyph embedding for Graph similarity.

4.3 Radical LCS Similarity

RLCSSim aims at measuring the similarity between radical systems. Here we represent a character as a radicals sequence and use longest common subsequence (LCS) to measure glyph similarity of characters. In this paper, each glyph is represented as a sequence of their smallest unit of radicals: $Seq(r_1, r_2, \ldots r_k)$. k is the number of radicals of that glyph. The sort order of the radicals r is determined by their positions within the character, which follows the rules of first left, then right; first up, then down; and first inside, then outside.

Eq.4 shows the RLCSSim between glyphs: RLCS means the longest common subsequence of same or related radicals between Seq_i and Seq_j. When calculating RLCS, we not only consider the same radical pairs, but also related radical pairs in derivative, indicative, or universal aspects (Section 4.1). If the corresponding two radicals are the same one, the RLCS will add 1, whereas if the two radicals are related, the RLCS will add a hyper-parameter $\theta, 0 < \theta < 1$. After getting the similarity of glyphs, the similarity between characters can be acquired according to Eq.3.

$$RLCSSim(g_i, g_j) = \frac{2 \times |RLCS(Seq_i, Seq_j)|}{|Seq_i| + |Seq_j|} \tag{4}$$

4.4 Graph Similarity

RLCSSim is discrete, and only covers character pairs sharing related radicals. In order to represent glyphs in high-dimensional vectors, and to acquire similarities among all character pairs, we introduces GraphSim. Here we construct an undirected graph $Graph$ based on ZiNet with the purpose of associating all Chinese glyphs through radicals. As shown in Figure 4, the set of nodes N includes character c, glyph g and radical r. There are three types of relations in $Graph$: $R_1(c, g)$, $R_2(g, r)$ $R_3(r, r)$: R_1 describes the inclusion relationship between characters and glyphs; R_2 describes the inclusion relationship between glyphs and radicals; R_3 contains derivative, indicative and universal relationships (Section 4.1) between radicals.

As the next step, based on the $Graph$, the random walk algorithm node2vec (Grover and Leskovec, 2016) is used to generate glyph embedding $\vec{G}$ of glyph nodes, while cosine similarity is used to obtain the similarity between glyphs. Finally, the GraphSim between characters can be acquired as the same way of Eq.3.

5 Evaluation

5.1 Design of Evaluation

We used Oracle data as the sample for evaluation, which contains 2543 Oracle characters, 2912 glyphs, 586 radicals and 15,175 character images; among them, 1283 characters are undeciphered up to now. The characters meanings cover each domain in that ancient age.

Experts were invited to further manually annotate the dataset: (1) There were 5400 Oracle character pairs randomly selected from the 2543 characters. Experts were asked to score them regarding glyph similarity. The corresponding value ranges from 0 to 10; the most similar character pair should be scored as 10. Three experts participated in this work, we selected the median as the final score for each pair of characters. (2) Experts were asked to provide less than five most similar characters to each Oracle character in sample. One expert firstly annotated similar characters. Then, another expert gave verification and deleted incorrect characters he thought. Finally, we got a total of 6405 similar pairs; on average, 2.5 similar characters were provided for each Oracle character, which have been represented as: $HSimSet\{(c_1, c_{11}), \cdots, (c_i, c_{in}), \cdots\}, i \leq 2543, n \leq 5$.

There are three quantitative and qualitative evaluation indicators:

- **Correlation:** Spearman's correlation was used to evaluate the correlation between similarity scores annotated by experts and our methods in 5400 pairs of Oracle characters.

Method	Top-5	Top-10	Top-20	Top-50	Top-100	Top-200
PicSim	19.53%	24.03%	29.74%	41.25%	50.27%	59.25%
RLCSSim	52.63%	65.21%	74.91%	86.15%	91.83%	95.93%
GraphSim	53.90%	64.84%	74.96%	85.92%	91.69%	96.03%
RLCSSim+PicSim	42.39%	52.51%	64.59%	78.61%	87.63%	94.53%
RLCSSim+GraphSim	**59.75%**	**70.37%**	78.86%	88.70%	93.99%	97.38%
RLCSSim+PicSim+GraphSim	57.13%	69.49%	**79.75%**	**89.41%**	**95.08%**	**97.86%**

Table 3: Results of coverage of the six methods in Top5–Top200 recommendations (the recommended size k was set to 5-200 according to the application scenarios in researches).

The value would be closer to 1 if it shows stronger positive correlation, conversely, it would be closer to -1.

- **Coverage:** The proportion of the 6405 similar character pairs appearing in the top-k similar character pairs provided by our methods (Eq.5), where the indicator aims at evaluating how much information that users need to browse to get the relevant one. $MSimSet\{(c_1, c_{11}), \cdots, (c_i, c_{ik}), \cdots\}, i \leq 2543, k \leq 2543$ to represent the top-k set of character pairs given by our methods.

$$Coverage = \frac{|HSimSet \cap MSimSet|}{|HSimSet|} \quad (5)$$

- **Qualitative analysis:** We show the top-5 recommendation examples to evaluate the performance and show potential semantic relations at radical level captured by the method.

5.2 Configuration

In the experiment, the number of layers of the ResNet network was 18, batch size was 64 and the learning rate was 0.001. The network was trained through 90 epochs. The hyper-parameter α was 0.4 and θ in RLCSSim was 0.7. The node2vec algorithm for GraphSim was implemented by using the OpenNE[1] tool; the dimension of the output glyph vector was 50.

In addition, this paper designed three combinations of basic methods: RLCSSim+PicSim, RLCSSim+GraphSim, and RLCSSim+PicSim+GraphSim. Their scores were set to the weights of basic methods. For the first two combinations, the weights of each basic method were 0.5 in both. Regarding RLCSSim+PicSim+GraphSim, the weights were respectively set to 0.4, 0.3 and 0.4. We extra annotated 100 ancient character pairs to set these hyper-parameters. The code of experiment can be acquired here[2].

Method	Correlation	
	score	p-value
PicSim	0.3241	<.001
RLCSSim	0.8188	<.001
GraphSim	0.7763	<.001
RLCSSim +PicSim	0.7614	<.001
RLCSSim+GraphSim	0.8391	<.001
RLCSSim+PicSim+Graph	**0.8422**	<.001

Table 2: Results of Spearman's correlation.

5.3 Results and Discussions

Spearman's correlations regarding the six methods are shown in Table 2; all of them show positive correlations respect to scores from experts. More in detail, RLCSSim+PicSim+GraphSim has the strongest positive correlation, corresponding to 0.8422, while the performance of PicSim method is not as good, with a 0.3241 value.

Table 3 shows the results of coverage indicator. RLCSSim+GraphSim achieved the best performance in the top5 and top10 recommendations, while, when dealing with larger recommendations size (top20 - top200), the effect of RLCSSim+PicSim+GraphSim has the most positive outcome. In a top-5 recommendation, four methods cover more than half of similar characters, while in top-200 recommendation, the coverage enhanced to more than 97% for RLCSSim+PicSim+GraphSim.

As results show, RLCSSim and GraphSim that are powered by knowledge of radical systems perform better than PicSim both in terms of corre-

[1] https://github.com/hengdos/OpenNE

[2] https://github.com/YangChiJLU/AncientChineseCharSim

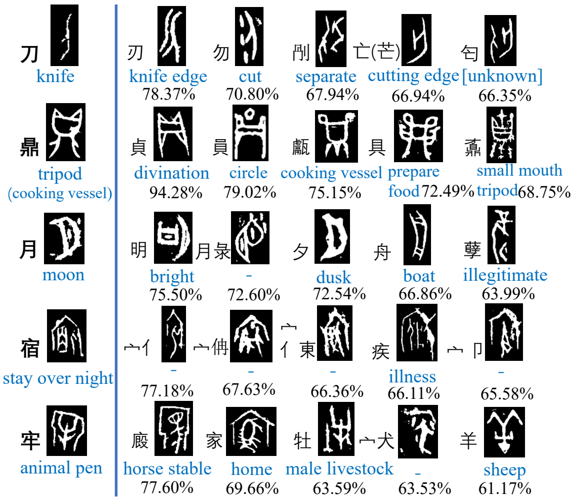

Figure 5: Cases of top-5 characters of glyph similarity (for each character, the Image, Title (e.g., "刀") in modern Chinese, English Annotation (e.g., "knife") and the Similarity Score (e.g., 78.37%) are shown. If the character is undeciphered, the English Annotation of it is "-" and the Title is written as the combination of the Titles of its Radicals (e.g., "宀亻" (house;person)).

lation and coverage. PicSim is suitable to comparing similarity between shapes of single pictographic characters. However, though some of character pairs show similar shapes, they are not similar at radical systems level. Thus, PicSim reduced the coverage of RLCSSim+PicSim, and RLCSSim+PicSim+GraphSim, in the case of small recommendations size. However, PicSim is meaningful to discover new similarities as the supplements of knowledge-powered methods. In larger size of scenarios, RLCSSim+PicSim+GraphSim performs better than only RLCSSim+GraphSim. Overall, the results show that radical systems are the crucial indicator for glyph similarity considered by human experts. It is necessary to represent, and calculate the potential relationships between radical systems of character pairs, rather than only consider characters as pictures. In application scenarios of small recommendations size, RLCSSim + GraphSim can be the best choice. In larger size of recommendations scenario, a combination of the three methods is the best choice.

5.4 Qualitative Analysis

Figure 5 shows five top-5 recommendations of the RLCSSim+PicSim+GraphSim method; the first three examples are single pictographic characters, while the other two are instances of compound characters which are formed with more than one radical. From the examples it can show many glyph similar character pairs are also related in semantics.

The first reason is that glyph similar pictographic characters always semantic related, which can be captured by PicSim method. As the figure shows, similar characters of "刀" (knife) are related to the knife edge, and the cut behavior, while similar characters of "鼎" (tripod) are mostly related to vessels for sacrifices and food. These characters have similar shapes, thus they can be recognized by PicSim. Another significant reason is that our method is also knowledge-powered, which can capture potential relations at the radical level. Regarding the compound character "宿" (stay overnight), at the radical level, all of meanings of recommended characters deal with a person doing activities in the house. Analogously, "牢" (animal pen) is formed by radicals "宀" (house) and "牛" (cow), and three similar characters are also combined by animals and houses: for instance, the most similar character "廏" (horse stable) is formed by "宀" (house) and "馬" (horse), whose meaning is also related to "animal pen". The character gets higher similar score because RLCSSim and GraphSim captured the semantic similarity between the radicals of "馬" (horse) and "牛" (cow).

In addition, this method is inclined to give higher scores for character pairs with potential relations. For instance, regarding the recommendations of character "月" (moon), the recommended character "夕"(dust) and "月" (moon) were derived from the same character. And another recommended character "舟"(boat) is the diachronic substitutable radical of "月": some Chinese characters (e.g., "前" (to forward)) were formed by "舟" (boat) in ancient age; however, today, their radicals have been changed to "月" (moon).

6 Implications and Future Work

This work firstly put forward a diachronic Chinese lexical resource, which expanded the architecture of Princeton WordNet by adding several layers to describe diachronic characters under the lexical layer. Word was regarded as the basic unit in most existing semantic lexical databases (George, 1995; Bond and Foster, 2013; Navigli and Ponzetto, 2012; Giunchiglia et al., 2017); however, based on our investigations, glyphs or radicals of Chinese characters can also show semantics, which have been used to enrich input information in several NLP tasks (Meng et al., 2019; Tao et al., 2019; Sun et al., 2021; Tao et al., 2021). Besides, in interdisciplinary researches with historical linguistics, Chinese history and paleography, etc., diachronic characters and words, glyphs and semantics were always discussed together because of their close links, while in the low-resources background, existing NLP algorithms have not been widely applied in these fields.

The significance of ZiNet is to give a more complete architecture to support diverse NLP tasks: it introduces not only lexical, but also glyph and character information, not only works for modern Chinese, but also ancient Chinese, or regarding them in the same diachronic space. We hope this work can enlighten diversity of the architecture of language resources and promote development of more NLP tasks in interdisciplinary researches.

At the application level, ZiNet hold potential for knowledge powered Chinese NLP and image processing algorithms, especially in interdisciplinary researches, such as cognate discovery, word sense tracking and rubbing character recognition. ZiNet can also support platforms and provide experts in related fields with domain knowledge and quick information suggestions. For instance, giving retrievals of the evolution timeline of characters and words, annotated unearth document corpus and recommendations of similar characters at various historical times.

In future work, ZiNet will be further expanded to other historical periods, and synsets will be linked into conceptual ontology layer to describe the topics of Chinese in varied historical periods. In application level, we will apply ZiNet in other knowledge powered tasks, for instance, using radical knowledge to enhance performance of ancient character image recognition. And we will further explore that how it can help research and decipher ancient characters. Meanwhile, we are developing a platform to support services of ZiNet, which will be open in near future.

7 Conclusion

This paper proposed ZiNet, a diachronic Chinese knowledge base, which is the first structured resource dedicated to describe the relations, and evolution of Chinese characters and words. Based on ZiNet, we demonstrated methods for calculating glyph similarity between ancient Chinese characters. Results show a strong positive correlation between the scores obtained from our method and from experts. We hope this work can serve experts in Chinese linguistics, history and related fields.

8 Ethics

Data of ZiNet was mainly from School of Archaeology, Jilin university, we got the permission for further development. Other data was processed from ancient dictionaries, which are open for access and researches. ZiNet also has limitations. Since the ancient characters were thousands of years away from now, lots of information was lost, and there are also disputes in existing academic theories, such as identity and meaning of a certain character, the character to which a glyph belongs, etc. As a result, inevitably, ZiNet is incomplete, and tends to be in line with the "mainstream" theories that is also possible to be incorrect proved in future. Therefore, in some cases, glyph similarity measurement and other applications based on ZiNet may produce misleading and omission. Relevant users can use ZiNet and applications to get suggestions efficiently; however, they need to rely on their own professional knowledge for judgment. All the same, we believe the positive impact of our work far outweighs the limitation.

Acknowledgements

This work is partially supported by National Natural Science Foundation of China through grants No.62077027.

References

Adam St Arnaud, David Beck, and Grzegorz Kondrak. 2017. Identifying cognate sets across dictionaries of related languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2519–2528.

Khuyagbaatar Batsuren, Gábor Bella, and Fausto Giunchiglia. 2020. Cognet: A large-scale cognate database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145.

William G. Boltz. 1986. Early chinese writing. *World Archaeology*, 17(3):420–436.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1352–1362.

Chinese Academy of Social Science (CASS). 1984. *Yin Zhou Jinwen Ji Cheng (Jinwen integration in Shang and Zhou Dynasty)*. Zhonghua Book Company, Beijing.

Miller A. George. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Fausto Giunchiglia, Khuyagbaatar Batsuren, and Gabor Bella. 2017. Understanding and exploiting language diversity. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pages 4009–4017.

Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 855–864.

Moruo Guo and Houxuan Hu. 1978. *Jiaguwen He Ji (The Comprehensive Dictionary of Oracle Characters)*. Zhonghua Book Company, Beijing.

Xu Han, Yuzhuo Bai, Keyue Qiu, Zhiyuan Liu, and Maosong Sun. 2020. Isobs: An information system for oracle bone script. In *Proceedings of the 2020 EMNLP*, pages 227–233.

Bradley Hauer and Grzegorz Kondrak. 2011. Clustering semantically equivalent words into cognate sets in multilingual lists. In *Proceedings of 5th international joint conference on natural language processing*, pages 865–873.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1063–6919.

Qingju Jiao, Yuanyuan Jin, Yongge Liu, Shengwei Han, Guoying Liu, Nan Wang, Bang Li, and Feng Gao. 2021. Module structure detection of oracle characters with similar semantics. *Alexandria Engineering Journal*, 60(5):4819–4828.

Jiaming Luo, Yuan Cao, and Regina Barzilay. 2019. Neural decipherment via minimum-cost flow: From ugaritic to linear b. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3146–3155.

Yuxian Meng, Wei Wu, Fei Wang, Xiaoya Li, Ping Nie, Fan Yin, Muyu Li, Qinghong Han, Xiaofei Sun, and Jiwei Li. 2019. Glyce: Glyph-vectors for chinese character representations. In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, pages 2742–2753.

Roberto Navigli and Simone P. Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Jane Qiu. 2014. Ancient times table hidden in chinese bamboo strips. *Nature*.

Xigui Qiu, Gilbert L. Mattos, and Jerry Norman. 2000. *Chinese Writing*. The Society for the Study of Early China and The Institute of East Asian Studies, University of California,Berkeley, California.

Edward L. Shaughnessy. 1991. *Sources of Western Zhou History: Inscribed Bronze Vessels*. University of California Press, Berkeley, Los Angeles, Oxford.

Shen Xu. 1963. *Shuo Wen Jie Zi*. Zhonghua Book Company, Beijing.

Benjamin Snyder, Regina Barzilay, and Kevin Knight. 2010. A statistical model for lost language decipherment. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1048–1057.

Zijun Sun, Xiaoya Li, Xiaofei Sun, Yuxian Meng, Xiang Ao, Qing He, Fei Wu, and Jiwei Li. 2021. Chinesebert: Chinese pretraining enhanced by glyph and pinyin information. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 2065–2075.

Hanqing Tao, Shiwei Tong, Kun Zhang, Tong Xu, Qi Liu, Enhong Chen, and Min Hou. 2021. Ideography leads us to the field of cognition: A radical-guided associative model for chinese text classification. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21)*, pages 13898–13906.

Hanqing Tao, Shiwei Tong, Hongke Zhao, Tong Xu, Binbin Jin, and Qi Liu. 2019. A radical-aware attention-based model for chinese text classification. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, pages 5125–5132.

Chongsheng Zhang, Ruixing Zong, Shuang Cao, Yi Men, and Bofeng Mo. 2020. Ai-powered oracle bone inscriptions recognition and fragments rejoining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*, pages 5309–5311.