JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 33, 1195-1211 (2017)
DOI: 10.6688/JISE.2017.33.5.6

Encoding and Ranking Similar Chinese Characters

MING LIU[1], VASILE RUS[3], QIANG LIAO[2] AND LI LIU[4]
[1]*School of Computer and Information Science*
[2]*School of Literature*
Southwest University
Chongqing, 400715 P.R. China
E-mail: {mingliu; liaoq}@swu.edu.cn
[3]*Department of Computer Science*
University of Memphis
Memphis, 38152 TN, USA
E-mail: vrus@memphis.edu
[4]*School of Software Engineering*
Chongqing University
Chongqing, 400044 P.R. China
E-mail: dcsliuli@cqu.edu.cn

Automatically detecting similar Chinese characters is useful in many areas, such as building intelligent authoring tools (*e.g.* automatic multiple choice question generation) in the area of computer assisted language learning. Previous work on the computation of Chinese character similarity focused on detecting character glyph similarity while ignored the importance of other character features, such as pronunciation and meaning. In this article, we present a way to encoding 4,500 simplified Chinese characters in terms of character glyph, pronunciation and meaning, annotating similar Chinese characters and automatically ranking similar characters based on the approach of learning to rank. The experiment results indicated that this approach could be useful for ranking and recognizing similar Chinese characters in terms of glyph, pinyin and semantic meaning. Moreover, it has been found that the learning to rank Listwise (ListNet) method was more effective than Pointwise (MART) and Pairwise (RankNet).

Keywords: natural language processing, Chinese character encoding, character similarity measurement, learning-to-rank, machine learning

1. INTRODUCTION

A Chinese sentence comprises of a sequence of words that are not separated by spaces. Normally, one Chinese word consists of a sequence of characters, although there are Chinese words that contain only one Chinese character. The meaning of a word is often highly related to the characters of which it is comprised. For example, "汽车" (automobile) and "火车" (train) are both kinds of "车" (vehicle). Moreover, each Chinese character is composed of one or more components. Usually, one component, the "semantic radical", carries the meaning of a character to a certain degree and the other, "phonetic radical", indicates the pronunciation. Characters sharing the same semantic radicals have related or similar meanings. For example, the characters "吃" (eat) and "喝" (drink) have the same leftmost component "口" (mouth). The shared aspect of meaning indicates that the mouth performs both actions. Meanwhile, characters sharing the same phonetic radicals have similar pronunciations. For instance, "漂" /Piao1/ (to flow) and "飘" /piao1

Received March 21, 2016; revised June 23 & August 10, 2016; accepted August 12, 2016.
Communicated by Chao-Lin Liu.

/ (to float) have similar pronunciations, since they share the same phonetic radical, "票" /piao4/. It is easy to decompose a word into characters and to decompose a character into components according to their forms. The semantic radicals are usually used as the index for words in a Chinese dictionary, and the Chinese government has published a national standard [1] to associate each character to a specific radical.

Characters that are similar in their appearances, pronunciations or meanings are useful for computer assisted language learning [2]. Multiple choice tests have proved to be an efficient tool for measuring students' achievement. When preparing multiple questions for assessing students' knowledge about words in a computer assisted environment, a teacher or pedagogical expert provides test sentences and indicates the target characters to be tested. The original sentences are then transformed into cloze questions by replacing the target character with a gap/blank and number of answer choices, one of which is the target character. Other choices called "distractors" as they are incorrect answers but related to the correct answer to various degrees. Generating plausible distractors is the key to an automatic multiple choice question generation systems [2], where distractors and the target character are confusing, which are similar in pronunciations or appearances. Thus, constructing and ranking confused distractor list is important for this application. In addition, Chinese students at primary schools are normally required to point out and correct "erroneous words" in test items, where an incorrect Chinese character are introduced intentionally when teachers prepare the test items. Creating incorrect character correction tests is a time-consuming task since the different incorrect characters are used in a test at a time. Thus, automatically ranking the list of similar incorrect characters is also useful in this type of application [3]. Moreover, the ranked lists of similar Chinese characters or confusion sets are useful in the area of Chinese spelling check [4, 5]. In fact, a series of Chinese spelling check shared tasks have been held in special interest group on computer-assisted language learning in 2013, 2014 and 2015.

The primary focus of our study is to build a model, which can automatically rank the list of candidate characters, based on the similarity between the target character and the candidate character. Ranking similar characters is particularly important in these applications because the same test items should not be repeatedly deliver to students. In other words, the distractors in a multiple choice question test or the incorrect characters in an incorrect character correction test should not be used repeatedly in every test item. If the list of similar candidate characters is ranked, the teacher can easily select several characters as distractors or incorrect characters from top ranked characters in that list to create test items at a time. Another importance of the ranking is that the ranked characters indicate the level of difficulty in a test item because that the top ranked characters are more confusing or similar to the target character than the low ranked characters. For example, Liu *et al.* [6] have built an environment for assisting the preparation of such test items, which do not repeatedly use the same incorrect characters at the same time by providing the list of ranked incorrect characters.

In psycholinguistics, several studies have examined how similar Chinese characters influence the judgments made by skilled readers of Chinese. Some researchers showed that stroke analysis and component decompositions were primary stages of printed character recognition [7]. The result indicated that orthographic similarity played a role in character recognition. Moreover, researchers [8, 9] focused on whether visual complexity and sub lexical phonology influenced character investigation. They found that the

phonology of the character contributed to character recognition. Other researchers [10] suggested that semantic radicals played a role in character recognition. Thus, these features contributing to character recognition have been considered in our proposed computational model.

The computation of Chinese character semantic and graphical similarity has attracted considerable attention in the field of natural language processing. Each Chinese character has its own meaning. Most of their single characters can be used as separate words. A number of language resources, such as HowNet [11], were developed to measure the semantic similarity of Chinese characters. On the other hand, researchers [3, 12-14] proposed computational methods for calculating the similarity of Chinese character glyph based on the character structure and its components' similarities. However, this algorithm requires human experts to manually provide similarity scores for the individual component of the Chinese characters and also assign weights to the different component comparison measures. Overall, in order to build a useful computational model for the character similarity measurement, we need to encode characters in a way that captures character features, such as glyph [12] and pronunciations [3]. To the best of our knowledge, there is no existing language resource, which is available to encode Chinese characters in a more comprehensive way and translate them into features used by statistical machine learning models.

In this article, we present a novel computational approach to measure the Chinese character similarity that considers different aspects of a character, including strokes, structure, Pinyin, semantic radical and semantic meaning. Specifically, the major contributions of this paper are the following:

- Encoded 4,500 simplified Chinese characters regarding to structure, semantic radical, Pinyin, stroke and meaning obtained from HowNet, and manually annotated the 520 lists of ranked characters generated from featured articles in Chinese elementary schools.
- Applied the approach of learning to rank to the task of ranking similar Chinese characters based on the character glyph, pronunciation and semantic meaning. Specifically, we evaluated three types of learning to rank algorithms including MART, RankNet and ListNet.
- Deeply analyzed the importance of features used in the ranking model. The experiment result showed that these features were significantly correlated to the similarity score.

The remainder of this paper is organized as follows: Section 2 describes the relevant work relating to the computation approach to Chinese character similarity measurement. Section 3 and 4 describe our character-encoding schema, data collection and annotation process. In Section 5 and 6, three experiments and their results are described. The paper concludes in Section 7 with a discussion of the overall approach as well as lines for future exploration.

2. BACKGROUND

2.1 Computational Approaches to the Similarity Measurement of Chinese Character

Chinese computational linguistics focused on measuring the similarity of Chinese

character glyph. Liu and Li [12] proposed methods for identifying visually similar Chinese characters by adopting and extending the basic concepts of the Cangjie method [13]. The Cangjie method defined 24 basic elements in Chinese characters and a set of rules to decompose Chinese characters into these basic elements. They extended the Chinese character encoding method by adding graphical structure information [14]. However, they did not clearly propose the similarity measurement algorithm based on the encoding method. Based on the extended Cangjie encodes, Liu *et al.* [3] proposed three similarity measurement algorithms. The first algorithm simply calculates the total number of matched elements between two characters. The second algorithm considers the structure and the location of the matched elements. One point is added if the structure is matched or the location of the shared elements is matched. The third algorithm computes the similarity in three steps. First, a character is concatenated into the parts of a Cangjie code. Then, the longest common subsequence (LCS) of the concatenated codes of the two characters being compared is computed. Lastly, a Dice coefficient is used to calculate the similarity. The study results showed that the third algorithm significantly improved the inclusion rates for visually related errors, where more actual incorrect characters are identified. But, as the authors noted, the parameters used in the similarity scoring function were not scientifically chosen, but were selected heuristically.

Similarly, Song *et al.* [15] considered the structure information and proposed a similarity measurement algorithm for Chinese characters. This algorithm first iteratively decomposed a compound Chinese character into a smaller component and gave a similarity score between compared smaller components. The final score is then calculated by summing up the weighted similarity scores between smaller components in iteration. The weight reflects on the level of the similarity between smaller components. Experts have predefined three levels from low to high. However, the problem with this approach is that the system performance can be influenced by the weights defined by human and the evaluation is unclear.

Liu *et al.* proposed [3] four categories of phonological similarity between two characters: same tone (SS), same sound and different tone (SD), similar sound and sound and same tone (MS), and similar sound and different tone (MD). Their system can select a list of phonologically similar characters for a given character under one of the four categories. The aim of their study is to produce incorrect character lists for generating incorrect character correction tests. The experimental results show that SS can capture better than 89.8% of the phonologically similar incorrect characters by an average of 12.2 characters. In addition, Chang [16] defined some simple rules to measure the phonological similarity between two Chinese characters. These rules compare two characters with their Mandarin phonetic symbols of its initial, medial, final and tone respectively to measure phonetic similarity. However, the evaluation results of the phonological similarity were not reported.

2.2 Chinese Lexicon Knowledge Base

Since 1980s, Efforts has been made in building Chinese semantic lexicon, such as, a dictionary of synonyms [17] and Chinese orthography database [18]. The problem with the existing language resources is that none of these language resources are organized on the basis of lexical taxonomical semantic relationships. However, with the release of

HowNet, a large vocabulary bilingual general knowledge base, research on HowNet based semantic similarity computation methods becomes available and made some progress [11].

HowNet is a knowledge base unveiling inter-conceptual relationships and inter-attribute relationships of concepts as connoting in lexicons of the Chinese and their English equivalents [11]. Unlike WordNet or Chinese Thesaurus of Synonyms [17], HowNet does not simply connect all the concepts in a hierarchical tree structure. It focuses on using sememes to describe concepts defined in a language called Knowledge Database Markup Language. Sememes are regarded as the basic unit of the meaning. HowNet contains 55501 concepts and 1620 sememes, 7152 Chinese characters and 92159 Chinese words. In HowNet, word similarity is measured by first calculating the sememe similarity, then concept similarity and finally the word similarity. The sememe similarity measure the semantic distance between two sememe nodes in a hierarchical tree structure [19]. In our study, we used the HowNet to obtain the semantic similarity between two characters as a feature used for the similarity measurement model.

2.3 Learning to Rank

Learning to rank is a relatively new research area, which received increasing attention in both the Information Retrieval and Machine Learning research communities, during the past decade. Most of approaches to learning to rank are designed as supervised machine learning approaches, *i.e.* learning a target concept from expert labeled instances. Typically, instances are assigned a (binary or ordinal) score or label indicating their relevance to the target concept as decided by an independent, expert judgment. In the training phase, a ranking function is learned based on a set of features the expert labels. In the testing phase, the ranking function is used to rank a new set of instances and generate a ranked order.

According to how they treat the sets of ratings and loss functions used during training, Cao *et al.* [20] classify learning to rank approaches into 3 categories: (1) Pointwise Approach: learning a classifier or regression model. These methods assume that each query document pair has either (a) a numerical or ordinal (rank) score associated with it or (b) a relevance label in one of two or more classes; (2) Pairwise Approach: learning the Pairwise preference of candidate documents rather than their absolute rank. The goal is now to learn a binary classifier that minimizes the number of incorrectly ordered pairs; (3) Listwise Approach: optimizing the loss function for ordering a set of candidate documents. Unlike Pointwise and Pairwise methods where a loss based on the rank of the individual candidate answer or of a pair is minimized; in Listwise methods, a direct loss (an appropriate evaluation measure defined by the user) is minimized between the true ranks of the list and the estimated ranks of the list.

Various machine-learning algorithms have been used to implement these approaches. For the Pointwise approach, the objective in the regression model is to find a model that predicts this score correctly through ordinal regression methods such as Multiple Additive Regression Tree (MART) [21]. The Pairwise approaches are considered more effective than Pointwise approaches because pairs of document instances are considered. The algorithms used in Pairwise approaches include RankNet [22]. Listwise approaches are more recent developments and have been shown to reach scores similar to or better

than the other two types of methods for information retrieval tasks [28]. Examples of Listwise methods include ListNet [20].

The general idea of ranking the output of a system using learning to rank approach has been explored in sentence parsing, natural language generation and dialogue systems. For example, Collins and Koo [23] presented methods for reranking syntactic parse trees from a generative parsing model using a discriminative ranker that can consider complex syntactic features. In our study, we applied and evaluated learning to rank (MART, RankNet and ListNet) in the Chinese character similarity computation.

3. METHOD

3.1 Chinese Character Encoding

We have selected 4,500 most common Chinese characters established by State Council of the People's Republic of China for encoding because these characters are suitable for beginner or intermediate learners to learn Chinese. Structure, semantic radical, strokes, pinyin and semantic meaning are the important characteristics in a Chinese character, and we have encoded characters in these aspects.

Professor Fu Yonghe in Beijing University identified thirteen kinds of Chinese character structures, which has been acknowledged by GB180302000 standard and ISO/IEC16046 standard. The structures are: left-right, left-middle-right, up-down, up-middle-down, full-round, up-three-round, left-three-round, down-three-round, up-left-round, down-left-round, up-right-round and symmetry. Linguistic postgraduate students manually annotated the structure based on the Xinhua Dictionary.

As we mentioned in the introduction section, a Chinese character can be decomposed into components called radicals, which is often a semantic indicator. There are 189 semantic radicals included in Xinhua Dictionary, such as 脚 (legs), 几 (table), 刀 (knife), 厂 (cliff) and 土 (earth).

A character glyph is based on the sequence of strokes. According to National Language Commission of China [24], five types of strokes are defined: (1) horizontal stroke; (2) vertical stroke; (3) left-falling stroke; (4) right-falling stroke; and (5) turning stroke. The stroke sequence was obtained from Data tang (http: //factory.datatang.com/en/index.html). Datatang is a professional data processing company, engaging in data collection and annotation.

Pinyin is the Chinese official phonetic system, which transcribes the pronunciations of Chinese characters into the Latin alphabet. Like English, pinyin uses 26 Latin letters to represent a character's pronunciation with the exception of "ü" standing for "v". The pinyin is also obtained from Datatang. However, it does not include other phonological elements, such as, onset, rime, and tone.

Each Chinese character usually has one or more meanings. The semantic meaning can be obtained from HowNet based on the formulas described in the section 2.2.

Table 1 shows an example of four encoded characters. These characters have the same structure and radical, slight differences in pronunciation and strokes and big differences in meaning.

Table 1. Examples of encoded Chinese characters.

Character	Structure	Semantic radical	Stroke	Pinyin	Meaning
辩	Left-middle-right	辛	4143113454143112	bian	Argue; debate
辨	Left-middle-right	辛	4143113434143112	bian	Distinguish; recognize
辫	Left-middle-right	辛	41431135514143112	bian	braid; pigtail
瓣	Left-middle-right	辛	4143113335444143112	ban	petal; segment;

3.2 Feature Definition

The features used in the ranking model were inspired from psycholinguistics research [8, 9] in Chinese character recognition mentioned in section 1. Therefore, these features should indicate the likelihood of generating the most similar Chinese characters in terms of the character glyph (stroke and structure), the pronunciation (pinyin), and semantic meaning (HowNet Similarity). Currently, there are 5 features defined as follows:

Structure: this Boolean feature indicates if two characters have the same structure.
Semantic Radical: this Boolean feature shows if two characters have the same semantic radical.
Stroke: this numeric feature indicates the difference in strokes between two Chinese characters. One Chinese character is represented a series of strokes in order. The difference is measured by the Levenshtein distance [25], where the minimum number of single stroke edits (*i.e.* insertions, deletions or substitutions) required to change one character into the other.
Pinyin: this numeric features also shows Levenshtein distance of Pinyin between two Chinese characters.
Semantic Distance: this numeric features describe the semantic similarity between two characters, which is measured by HowNet [19]. As we described before, the semantic similarity of characters is calculated based on the maximum similarity of concepts computed by the distance between sememe nodes in a hierarchical tree structure.

3.3 Data Collection and Annotation

The first dataset was created based on a list of vocabularies used in the textbook of elementary school. There were 35 Chinese articles randomly selected from grade 15 textbooks published by People's Educational Press. These articles were commonly used in elementary schools in China. Each article contained a list of new vocabularies (target characters) for students to learn. We used each character as a target character in the list to generate five candidate characters. The candidate characters were selected from our database containing 4500 encoded characters described in the previous section. The strategy of candidate character selection is that the candidate characters have the same structure, radical as the target character has. If there is no candidate character, which meets the requirement, we randomly selected candidate characters.

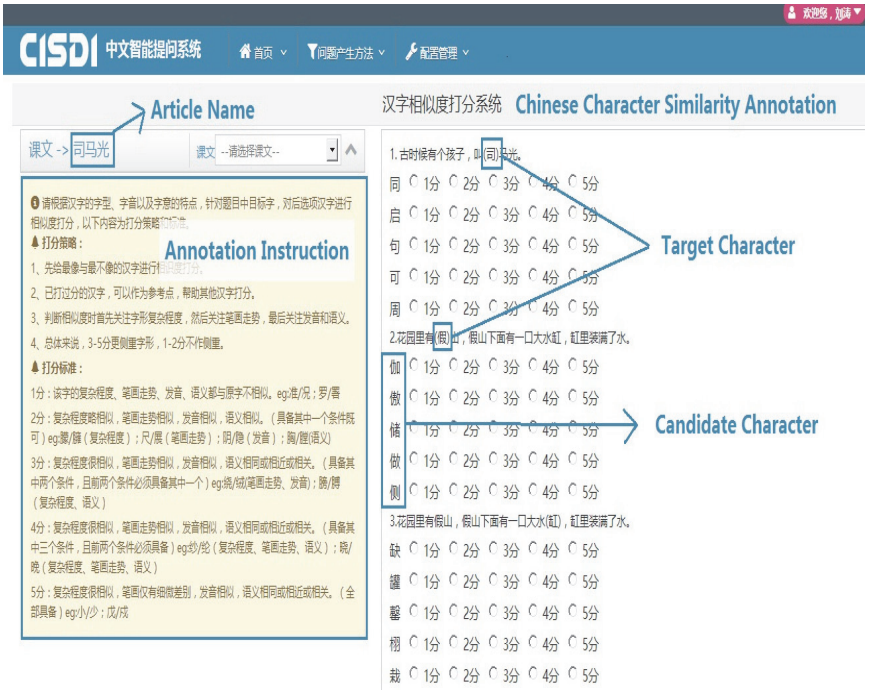

Fig. 1. A screenshot of the user interface to Chinese character similarity annotation.

Fig. 1 shows the web interface of Chinese character similarity annotation. Each annotator first reads and understands the annotation instruction including the annotation schema and strategy shown in the left side of the Figure and then selects article, reads the sentence containing the target character, and then rates the similarity between the target character and the candidate character.

The annotation procedure was the following: annotators first chose candidate characters, which are easy to rate (identified the most and least similar candidate characters), and then rated the confusing candidate characters. When it was difficult to assign scores for the candidate characters, annotators could use the rated candidate characters as a reference, such that similar characters about the same target character would receive similar ratings, and provide a more reliable gold standard. Our annotation schema for measuring character similarity is shown below. In general, mark 3 and 5 require that two characters should be at least similar in glyph, such as stroke trend and complexity

1 point: two characters do not share any of the features, such as similar complexity (the number of strokes and component structure), stroke trend, pronunciations or semantic meaning.

2 points: two characters share at least one of the features, similar complexity, stroke trend, pronunciations or semantic meaning. *E.g.* 朦/臃 (complexity); 尺/展 (stroke trend); 阴/隐 (pronunciation); 胸/膛 (semantic meaning)

3 points: two characters share at least two of the features, similar complexity, stroke trend, pronunciations or semantic meaning, including at least complexity feature or

stroke trend feature. *E.g.* 绕/绒 (stroke trend & pronunciation); 膀/膊 (complexity & semantic meaning)
4 points: two characters share at least three of the features, similar complexity, stroke trend, pronunciations or semantic meaning, including at least complexity feature and stroke trend feature. *E.g.* 纱/纶 (complexity, stroke trend, semantic meaning); 晓/晚 (complexity, stroke trend, semantic meaning)
5 points: two characters share four of the features, similar complexity, stroke trend, pronunciations or semantic meaning. *E.g.* 小/少; 戊/戍

Since these 35 articles contain 520 target characters (new vocabularies in each article), 520 lists of characters generated and each list contains a target character and five candidate characters. For each list, two Chinese linguistic postgraduate annotators independently rated the similarity between the target character and the candidate character based on the annotation schema. As a result, 520 ranked lists of characters were constructed. If there was a disagreement on rating, the third annotators joined the discussion. A pair of characters was finally scored only if a majority of the three raters agreed on the score. An inter-rater agreement of $r = .718$ (Pearson correlation) was obtained from the datasets acceptability ratings. This value corresponds to "high agreement". The major reason for causing some disagreements was that the similarity of the character complexity was difficult to decide. Besides, the pronunciation similarity could be defined more specific since it depends on three of the phonological elements – onset, rime, and tone. Within these 2,600 pairs of characters, we removed 14 pairs of characters since the semantic similarity of these pairs is not available in HowNet. Therefore, 517 lists of characters were used for this study.

The second dataset was constructed using a well-known Chinese language learning resource created by Jiang [26]. He identified 100 pairs of most confusing Chinese characters after analyzing more than 400 Chinese examination papers for primary and middle school students. In this study, we randomly generate a list of 9 candidate characters for a target character selected from a pair, which means that a list contains 10 pair of characters. Among them, one pair is from Jiang's dataset. Mathematically, the ranking value of this pair is 1 (relevant); otherwise, the rest of pairs in the list are 0 (irrelevant). Because 18 pairs of the Jiang's dataset contained characters that are neither in our database nor in the HowNet, we only used 82 pairs to generate the lists. Therefore, 82 lists of characters were used for this study.

4. EVALUATION MEASURES

The output of the model is a ranked list of candidate characters for each target character. The normalized discounted cumulative gain (*NDCG*) is the most popular measures for evaluating the performance of a recommendation system [27]. *NDCG@K* looks at the TOP *K* candidate characters, and assigns a higher weight to a very similar character that is ranked higher than one that is ranked lower. It varies from 0.0 to 1.0, with 1.0 representing the ideal ranking of the entities. *NDCG@K* is computed as

$$NDCG_k = \frac{DCG_k}{IDCG_k}. \quad (1)$$

Where $IDCG_k$ is the ideal discounted cumulative gain (*DCG*) when entities are sorted by relevance, it is to producing the maximum possible *DCG*. The *DCG* shows that highly relevant entities appearing lower in a search result list should be penalized as the graded relevance value is decreased logarithmically proportional to the position of the result. *DCG* accumulated at a particular rank position p is defined below:

$$DCG_k = rel_i + \sum_{i=2}^{k} \frac{rel_i}{\log_2 i}. \tag{2}$$

Where REL_i is the graded relevance of the result at position *i*. In our study, we used NDCG to measure the performance of learning to rank methods on the first dataset. In addition, we used other two common metrics, precision and reciprocal rank, to evaluate the learning to rank methods on the second dataset since the rank was induced by giving a binary judgment.

The mean reciprocal rank (*MRR*) is the average of the reciprocal of the rank at which the first correct answer is found. The equation is defined as below:

$$MRR = \frac{1}{T} \sum_{i=1}^{T} \frac{1}{Rank_i} \tag{3}$$

where $Rank_i$ means the rank of the relevant candidate character in a list. *T* refers to the number items in the list. *P* and *MRR* can also be evaluated at a given cutoff rank, considering only the topmost results returned by the system. This measure is called precision at *k* or *P@K* or reciprocal rank at *k* or *MRR@K*.

5. EXPERIMENTS

This section describes three experiments we conducted to evaluate the proposed approach to rank the Chinese characters. We evaluated the performance of the different learning to rank methods on the dataset 1 in the first experiment and on the dataset 2 in the second experiment. Specifically, three common learning to rank methods were used, MART (Pointwise), RankNet (Pairwise) and ListNet (Listwise), which are implemented in RankLib [28]. For all rankers, we used the default hyper-parameters suggested by authors, without any tuning.

5.1 Experiment on the First Dataset

We used 10 fold cross-validation to evaluate the performance of the three learning to rank methods on the first dataset. As we expected, the Listwise approach has superior performance in ranking. Table 2 shows that ListNet outperformed RankNet and MART in terms of all measures. We ran an ANOVA (Analysis of Variance) test on the performance result of these three methods with 95% confidence level and found significant results for NDCG@1 and NDCG@2. Then, we conducted a post hoc analysis with the Fisher's least significant difference (LSD) and found that the differences between ListNet and RankNet are significant ($p < 0.05$). The main reason is that the loss function

used in the ListNet can more properly represent the performance measures NDCG. This finding is in consistent with Cao *et al.*'s study [20].

We also investigated the impact of different feature sets on the ranking results and defined three feature models. Orthographic model: it contains character structure, strokes and semantic radical features; Phonological model: it includes pinyin feature; Semantic model: it covers semantic distance and semantic radical features. We selected the ListNet algorithm to implement the ranking model with different feature sets and evaluated them in the first dataset using ten fold cross-validation. Table 3 shows that the performance of ranking model with all features and orthographic features outscore those with phonological and semantic features across all the measures. The ANOVA tests showed significant differences in the performance (NDCG@1 and NDCG@2) of the ranking model with different feature sets. The LSD results implied that the ranking model with all features and orthographic features significantly outperformed the model with phonological features across NDCG@1 and NDCG@2, and the model with semantic features in NDCG @2. No significant differences are found between the ranking model with all features and the model with orthographical features across five measures. These results indicated that the orthographic features are generally more useful while the phonological feature is less effective because only pinyin is used in the phonological feature.

Table 2. The average scores of three learning to rank methods with all five features using ten fold cross-validation on the first dataset.

	MART	RankNet	ListNet
NDCG@1	0.644	0.569	0.687*
NDCG@2	0.708	0.648	0.756*
NDCG@3	0.759	0.732	0.805
NDCG@4	0.816	0.791	0.854
NDCG@5	0.843	0.808	0.866

* means the significant level at 0.05.

Table 3. The average scores of the ListNet algorithm with different feature sets using ten fold cross-validation on the first dataset.

	Orthographic Features	Phonological Feature	Semantic Features	All Features
NDCG@1	0.672	0.532	0.618	0.687*
NDCG@2	0.718	0.589	0.632	0.756*
NDCG@3	0.797	0.722	0.760	0.805
NDCG@4	0.823	0.756	0.809	0.854
NDCG@5	0.838	0.802	0.815	0.866

* means the significant level at 0.05.

5.2 Experiments on Ranking Confusing Chinese Characters

We first used the 517 lists of characters to train the ranking model optimizing NDCG@1, and then we applied this model to rank the 82 lists of characters in the second dataset. Table 4 shows that the ListNet outscored the other two algorithms across all measures. RankNet and MART reached similar scores. In addition, the result showed

that ListNet got 0.679 at P@1, which indicated that more than half of the most confusing character pairs from Jiang's study had been identified at rank 1. The result of 0.728 at MRR@5 implied that 93.8% (77 lists of confusing characters were recognized) of the most confusing character pairs had been recognized in top 5.

Table 5 shows that the candidate characters with higher ranks have the same structure, similar pinyin, same semantic radical or similar semantic meanings as the target characters have. For example, given the character 滔, 涛 got the highest rank since their structure, radical and pinyin are the same and their semantic meaning is similar (great wave). 沧 and 渔 take the second and third place respectively since they share the same structure and radical. 箭 has the lowest rank since it does not share only features, such as structure and pronunciation with 滔.

Table 4. The average scores of three methods on the second dataset.

Metrics	MART	RankNet	ListNet
P@1	0.482	0.346	0.679
P@5	0.151	0.163	0.187
P@10	0.1	0.1	0.1
MRR@5	0.593	0.580	0.728
MRR@10	0.616	0.625	0.773

Table 5. Examples of the system ranked candidate characters given target characters.

Target Character	Candidate Character List
滔	涛沧渔焕箭
弛	驰既成泻琅
渡	度渔立费是
惠	慧意厉度合

The list of candidate characters is ordered from left to right.

Table 6. The correlation between features and similarity scores.

	Meaning	Stroke	Radical	Structure	Pinyin
Pearson	.101**	.260**	.181**	.178**	.072*
Spearman	.092**	.269**	.186**	.182**	.043
Tau_b	.076**	.221**	.175**	.171**	.034

** indicates the significant level at 0.01 where * the significant level at 0.05.

5.3 Experiment on Correlation Analysis

In this experiment, we used the data analytic software SPSS to investigate the correlation between features and similarity scores given by human annotators based on the first dataset, which contains 2585 pairs of characters with similarity scores. Table 6 shows that four features including Meaning, Stroke, Radical and Structure separately significantly correlated to the similarity scores ($p < 0.01$, $n = 2585$) across three correlation coefficients. The pinyin feature significantly correlated to the similarity score based on the Pearson coefficient. In general, this result suggested that individual feature was significantly correlated to the similarity score.

6. DISCUSSION AND CONCLUSION

Automatically getting a list of ranked Chinese characters that are similar to a given character is a challenging task due to the complexity of Chinese characters. Most studies focused on computing the glyph similarity of Chinese characters and ignored the importance of pronunciation and semantic meaning features, although researchers in psycholinguistic [8, 10] found that phonology and meaning of the character also contributed to character recognition. In this study, we proposed a learning-to-rank approach to measure the similarity of characters regarding glyph, pronunciation and semantic meaning. The features used were inspired by the result in psycholinguistics that strokes, radical and semantic meanings are important for character recognition. In addition, we have constructed two datasets for evaluating our approach: the first dataset contains 520 lists of characters obtained from an elementary school textbook; the second contains 82 lists of characters constructed based on a well-known Chinese language learning resource created by Jiang [26].

In the first experiment, we evaluated the three different learning to rank methods in the first dataset. It has been found that the performance (NDCG) of the ListNet (Listwise) method outscored the MART (Pointwise) and ListNet (Pairwise) methods and yielded a score of 0.866 measured by NDCG@5 (1 is the ideal). This experiment demonstrated the Listwise approach has superior performance, which was also found in Cao *et al.*'s study [20].

In the second experiment, we used the three learning to rank models trained on the first dataset to rank 82 lists of characters in the second dataset. The study result showed that the ListNet could recognize 67.9% pairs of the most confusing characters at rank 1 and 93.8% in the top 5 results. This finding indicated that this model has potential in recognizing and generating confusing Chinese characters, which could be useful for automatic multiple question generation. Furthermore, the third experiment illustrated that the five features were significantly correlated to the similarity scores based on Pearson coefficient ($p < 0.05$), which indicated that the proposed features were useful.

Our current Chinese character encoding method is simple and easy to implement. The statistical ranking model naturally captures the similarity between two characters in appearance, pronunciation and semantic meaning based on the five features derived from the character encoding. In addition, the parameters used in the ranking model learned from the dataset rather than those parameters setup by their heuristics. Our experimental results suggested that the ranking model with all features were more effective than those with only phonological features or semantic features. However, this encoding approach has some limitations. One limitation is that some Chinese characters are not available in HowNet, which causes the problem of extracting the semantic distance feature of these characters. The semantic radical indicates the meaning of a Chinese character. Thus, one possible solution is that we can calculate the average similarity score between the target character and the list of candidate characters. This list is constructed by using the missing semantic meaning character's semantic radical information to select the candidate characters containing the same semantic radical. Another limitation is that current encoding method may not be suitable for measuring the similarity between traditional Chinese characters, particularly in orthographic similarity. The traditional Chinese character is more complex in glyph than the simplified Chinese character. Thus, the stroke feature is

not efficient to capture the similarity. For examples, simplified Chinese characters "儿" and "几" are visually similar, but their corresponding traditional characters "兒" and "幾" are significantly different. As the studies show [3], Cangjie codes are more efficient in measuring the similarity of traditional Chinese characters. Furthermore, there is still space for improvement on the computational model of character similarity measurement by adding more fine-grained features. For example, Pinyin is the only phonological feature used in the study. But, other phonological features, such as, onset, rime, and tone, were also important. In addition, the semantic distance between the context of the target character and the distractor could be a useful feature.

In sum, this article describes an approach to encode simplified Chinese characters, presents an annotation schema to rate the similarity of two Chinese characters, and apply learning-to-rank methods to build a ranking model, which can automatically rank the similar characters. In the future, we will focus on encoding more features into characters and integrating this model in our automated multiple choice question generation system and using it in a Chinese language-learning context at an elementary school. More specifically, we will investigate and evaluate different distractor generation strategies in a MCQ test. For example, candidate characters were first selected if they had the same structure, radical and similar frequency as the target character has. Then we can apply the ranking model (mixed strategy) to select the top n most similar candidate characters as distractor to be used in a MCQ test. Mitkov *et al.* [29] used the test item analysis procedure to evaluate different strategies, such as semantic similarity, distributional similarity, phonetic similarity to generate distractors. In addition, we will plan a similar user study to the one conducted by Aldabe and Maritxalar [30], who generated MCQs for a science vocabulary learning scenario in a Basque educational environment. Furthermore, we are planning to release this language resource containing the encoded simplified Chinese characters for enhancing the research community in the Datatang (http://factory.datatang.com/en/index.html).

ACKNOWLEDGMENT

This work is supported by National Natural Science Foundation of China (61502397), the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry (the 50th grants), and Fundamental Research Funds for the Central Universities under Grant No. XDJK2014A002, No. XDJK2017C024, No. SWU114005, CQU903005203326 and CQU0225005202017.

REFERENCES

1. Y. Chen, Y. Lin, J. Chen, and Y. Song, *Specification for Identifying Indexing Components of GB,13000.1 Chinese Characters Set*, Language and Literature Press, Beijing, 2009.
2. R. Mitkov, L. Anha, and N. Karamanis, "A computer-aided environment for generating multiple-choice test items," *Natural Language Engineering*, Vol. 12, 2006, pp. 177-194.

3. C.-L. Liu, M.-H. Lai, K.-W. Tien, Y. Chuang, S.-H. Wu, and C.-Y. Lee, "Visually and phonologically similar characters in incorrect Chinese words: Analyses, identification, and applications," *ACM Transactions on Asian Language Information Processing*, Vol. 10, 2011, pp. 1-39.
4. A. Carlson, J. Rosen, and D. Roth, "Scaling up context-sensitive text correction," in *Proceedings of the 13th Innovative Applications of Artificial Intelligence Conference*, 2001, pp. 45-50.
5. C.-J. Lin and W.-C. Chu, "A study on Chinese spelling check using confusion sets and *N*-gram statistics," *Computational Linguistics and Chinese Language Processing* Vol. 20, 2015, pp. 23-48.
6. C.-L. Liu, K.-W. Tien, Y.-H. Chuang, C.-B. Huang, and J.-Y. Weng, "Two applications of lexical information to computer-assisted item authoring for elementary Chinese," in *Proceedings of the 22nd International Conference on Industrial Enneering and Other Applications of Applied Intelligent Systems*, 2009, pp. 470-480.
7. M. Taft and X. Zhu, "Sub-morphemic processing in reading Chinese," *Journal of Experimental Psychology: Learning*, *Memory and Cognition*, Vol. 23, 1997, pp. 761-775.
8. C. A. Perfetti, S. Zhang, and I. Berent, "Reading in English and Chinese," *Orthography*, *Phonology*, *Morphology and Meaning*, R. Frost and L. Katz, eds., pp. 227-248, Amsterdam, 1992.
9. L. H. Tan, R. Hoosan, and W. T. Siok, "Activation phonological code before accessing to Chinese character meaning in written Chinese," *Journal of Experimental Psychology: Human Learning and Memory*, Vol. 3, 1996, pp. 621-630.
10. L. B. Feldman and W. W. T. Siok, "Semantic radicals contribute to the visual identification of Chinese characters," *Journal of Memory and Language*, 1999, pp. 559-576.
11. Z. Dong and Q. Dong, *HowNet and the Computation of Meaning*, World Scientific, Singapore, 2006.
12. C.-L. Liu and J.-H. Lin, "Using structural information for identifying similar Chinese characters," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, 2008, pp. 93-96.
13. B.-F. Chu, *Handbook of the Fifth Generation of the Cangjie Input Method*, [5 Dec. 2015, 2008].
14. D. Juang, J.-H. Wang, C.-Y. Lai, C. C. Hsieh, L.-F. Chien, and J.-M. Ho, "Resolving the unencoded character problem for Chinese digital libraries," in *Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries*, 2005, pp. 311-319.
15. R. Song, M. Lin, and S. Ge, "Similarity calculation of Chinese character glyph and its application in computer aided proofreading system," *Journal of Chinese Computer Systems*, Vol. 29, 2008, pp. 1964-1968.
16. T.-H. Chang, H.-C. Chen, and Y.-H. Tseng, "Automatic detection and correction for Chinese misspelled words using phonological and orthographic similarities," in *Proceedings of the 7th SIGHAN Workshop on Chinese Language Processing*, 2013, pp. 97-101.
17. J. Mei, *Chinese Thesaurus of Synonyms*, The Shanghai Foreign Language Audiovisual Publishing House, Shanghai, 1983.

18. H.-C. Chen, L.-Y. Chang, Y.-S. Chiou, Y.-T. Sung, and K.-E. Chang, "Chinese orthography database and its application in teaching Chinese characters," *Bulletin of Educational Psychology*, Vol. 43, 2011, pp. 269-290.
19. Q. Liu and S. J. Li, "Word similarity computing based on how-net," in *Proceedings of the 3rd Chinese Lexical Semantics Seminar Proceedings*, 2012, pp. 59-76.
20. Z. Cao, T. Qin, T. Y. Liu, M. F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *Proceedings of the 24th International Conference on Machine Learning*, 2007, pp. 129-136.
21. D. Cossock and T. Zhang, "Subset ranking using regression," *Learning Theory*, LNCS Vol. 4005, 2006, pp. 605-619.
22. C. Burges, T. Shaked, E. Renshaw, A. Lazier, and M. Deeds, N. Hamilton, G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the 22nd International Conference on Machine Learning*, 2005, pp. 89-96.
23. M. Collins and K. Koo, "Discriminative reranking for natural language parsing," *Computational Linguistics*, 2003, pp. 175-182.
24. *Standard Stroke Order of Commonly-Used Characters of Modern Chinese*, N. L. C., China, 1997.
25. G. Navarro, "A guided tour to approximate string matching," *ACM Computing Surveys*, Vol. 33, 2001, pp. 31-88.
26. Y. Jiang, "100 pairs of most confusing Chinese words," *Yu Wen Tian Di*, Vol. 7, 2006, pp. 19-20.
27. Y. Wang, L. Wang, Y. Li, D. He, W. Chen, and T.-Y. Liu, "A theoretical analysis of NDCG ranking measures," in *Proceedings of the 26th Annual Conference on Learning Theory*, 2013, pp. 1-33.
28. V. Dang, "Ranklib – a library of ranking algorithms," http: //www.cs.umass.edu/~ vdang/ranklib.html, 2015.
29. R. Mitkov, L. A. Ha, A. Varga, and L. Rello, "Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation," in *Proceedings of Workshop on Geometrical Models of Natural Language Semantics*, 2009, pp. 49-56.
30. I. Aldabe and M. Maritxalar, "Semantic similarity measures for the generation of science tests in basque," *IEEE Transactions on Leraning Technoigies*, Vol. 7, 2014, pp. 375-357.

Ming Liu (刘明) is an Associate Professor at the School of Computer and Information Science, Southwest University, China. He received the Ph.D. in Artificial Intelligence in Education at the School of Electrical and Information Engineering, The University of Sydney, Australia in 2012. His main research interests include natural language processing and learning analytics. He has been involved in research and development projects in the areas of computational linguistics in both Australia and China.

Vasile Rus is a Professor at the University of Memphis. Dr. Rus earned his Master of Science in Computer Science and Doctor of Philosophy in Computer Science degrees from Southern Methodist University at Dallas, Texas in May 1999 and May 2002, respectively. He has been involved in research and development projects in the areas of computational linguistics and information retrieval for more than 15 years.

Qiang Liao (廖强) currently is an Associate Professor of Literature in Southeast University. He received his BS from Sichuan Normal University in 1994, then his MS from Southwest Normal University in 2003, and then his Ph.D. from Nanjing Normal University, and then did his postdoctoral research in Beijing Normal University in 2012. His research interests include Chinese character encoding, classic Chinese, unearthed documents, and local culture.

Li Liu (刘礼) is an Associate Professor at Chongqing University. He had served as a Senior Research Fellow at the National University of Singapore. Li received his Ph.D. from the Université Paris-sud XI in 2008. His research interests are in pattern recognition, data analysis, and their applications on human behaviors.