Character Set Construction for Chinese Language Learning

Chak Yan Yeung, John Lee
Department of Linguistics and Translation
City University of Hong Kong
Hong Kong SAR, China
cyyeung91@gmail.com, jsylee@cityu.edu.hk

Abstract

To promote efficient learning of Chinese characters, pedagogical materials may present not only a single character, but a set of characters that are related in meaning and in written form. This paper investigates automatic construction of these character sets. The proposed model represents a character as averaged word vectors of common words containing the character. It then identifies sets of characters with high semantic similarity through clustering. Human evaluation shows that this representation outperforms direct use of character embeddings, and that the resulting character sets capture distinct semantic ranges.

1 Introduction

To promote efficient vocabulary acquisition, pedagogical materials may present the learner with a set of related words, rather than a single word. The set often consists of words belonging to the same "family"; for English, family members may share the same root, such as the words *special*, *specialize*, *specialty*, *especially*, etc. These families can be constructed in a straightforward manner from morphological databases and analyzers.[1]

An analogous strategy for teaching Chinese is the "character family", i.e. a set of characters that are similar in meaning and written form. A natural criterion for family membership is the semantic component, or semantic radical, of the character. The family based on the component 'sun', for example, includes the characters for 'sunny', 'sunshine' and 'dawn' as members (Table 1).

In comparison to word families in English, character families tend to exhibit less semantic regularity. Some family members may have unrelated meaning, or have obscured semantic relation in modern Chinese. For example, the character *cuò* 'wrong' belongs to the 'metal' family, and the character *zuó* 'past' belongs to the 'sun' family (Table 1). When preparing character sets for use in computer-assisted language learning (CALL) applications, manual selection is often necessary to ensure that the sets illustrate semantic regularity.

This paper investigates automatic construction of character sets. The proposed method can be expected to expedite the generation of these sets for more components and for a larger variety of semantic categories, with the goal of enhancing the coverage and effectiveness of CALL applications for learning Chinese.

2 Research Questions

For each semantic component, we define its "character family" to consist of all characters that contain the component. As shown in Table 1, not all family members have sufficiently related meaning to serve as good examples in pedagogical materials. Given a family, the *character set construction* task is to identify a subset of its characters that are semantically close. In designing an algorithm for this task, we address two research topics:

Character representation (Q1) The character representation should reflect the "overall" meaning of the character in a variety of contexts. We compare the use of character and word embeddings in constructing character sets (Section 4).

Subfamilies (Q2) All family members are traditionally viewed as capturing the general meaning of its semantic component. We investigate whether some families can be clustered into subfamilies to produce character sets with more tightly related meaning (Section 5).

[1] Examples include CELEX (Baayen et al., 1995) and Morfessor (Creutz and Lagus, 2006).

Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pages 59–63

Semantic component	Character family	
	Example character sets	Other characters
日 *rì* ‘sun’	晴 *qíng* ‘sunny’, 暉 *huī* ‘sunshine’, 曉 *xiǎo* ‘dawn’	昨 *zuó* ‘past’
金 *jīn* ‘metal’	銅 *tóng* ‘copper’, 鐵 *tiě* ‘iron’, 銀 *jín* ‘silver’	錯 *cuò* ‘wrong’
頁 *yè* ‘page’	頭 *tóu* ‘head’, 額 *é* ‘forehead’, 頸 *jǐng* ‘neck’	類 *lèi* ‘type’
女 *nǚ* ‘female’	Subfamily #1: 嫁 *jià* ‘marry’, 娶 *qǔ* ‘marry’, 婚 *hūn* ‘marriage’ Subfamily #2: 姨 *yí* ‘aunt’, 姐 *jiě* ‘older sister’, 妹 *mèi* ‘younger sister’	始 *shǐ* ‘begin’

Table 1: Each character family is associated with a semantic component and its members consist of all characters that contain the component. Semantic similarity can be strong for some family members (“Character sets” column) but less apparent for others (“Other characters” column).

3 Background

Chinese words are composed of characters. According to Li and Kang (1993), 81% of the characters are “semantic-phonetic compounds”, which can be decomposed into two components. The phonetic component gives pronunciation clues. The semantic component, often used for organizing characters into families (Table 1), indicates the semantic range of the character.

The rest of this section summarizes research on Chinese subword structures in CALL (Section 3.1) and in natural language processing (Section 3.2).

3.1 CALL for Chinese characters

There are considerable pedagogical benefits in highlighting the semantic regularity in character families (Tse et al., 2007; Leong et al., 2011). Many CALL applications for Chinese have therefore featured these families, including web-based tutorials (Chen et al., 2011) and Scrabble-like character formation games (Lam et al., 2001; Lee and Yeung, 2020).

It is however well known that not all members in a character family have related meaning in modern Chinese. A character is called *transparent* if its meaning is similar or directly related to that of its semantic component. For example, in the ‘metal’ family, the characters *tóng* ‘copper’ and *tiě* ‘iron’ are transparent, while the character *cuò* ‘wrong’ is not (Table 1). According to an analysis of primary school material (Chung and Leung, 2008), only 64% to 82% of the characters have meaning that is related or somewhat related to its semantic component. A direct consequence is that semantic components are not uniformly useful in aiding comprehension (Liow et al., 1999). Character sets therefore often require manual curation, which constrains their use in interactive CALL applications.

3.2 Subword representation in Chinese

Various algorithms have been proposed to train embeddings for Chinese at the subword level. Contextual embeddings such as BERT (Devlin et al., 2019) are designed to derive character embeddings in a specific sentential context. In contrast, identifying the general or overall meaning of a character is the main objective in the character set construction task. Our evaluation will focus on the use of context-free embeddings.

Context-free embeddings at the word-, character- and component levels (Lu et al., 2016; Yu et al., 2017; Cao et al., 2018; Devlin et al., 2019) have been applied to many downstream tasks in Chinese NLP, but there has not been any quantitative evaluation on their use in creating character sets for CALL. Besides direct use of character embeddings, a possible approach is to measure similarity between character and component embeddings, as suggested by a qualitative study on the ‘illness’ component (Yu et al., 2017). An alternative is to exploit embeddings of words formed by the character, although the character’s semantic contribution to different words may vary (Xu et al., 2016). Our study evaluates character sets produced by a number of these approaches.

4 Character representation

We address Q1 by evaluating two character representations for the character set construction task: given a family F, identify a subset of N characters, say $S = \{c_1, ..., c_N\} \subseteq F$, that have the most similar or related meaning.

4.1 Approach

A simple approach would be to retrieve characters with the closest meaning to the semantic component associated with F. This method can be problematic, however, since the dominant meaning of a component may differ from those of the family members. For example, members of the family associated with the component *yè* are semantically related to "head", but as a standalone character *yè* means 'page' in modern Chinese (Table 1).

We instead measure semantic similarity between characters. Using a set of 7.6 million sentences from Chinese Wikipedia, we trained context-free embeddings $\vec{c}$ and $\vec{w}$ for each character c and word w with the joint learning model proposed by Yu et al. (2017). We compare two methods for generating the representation $v(c)$ for a character c:

Character vector The baseline directly uses the character embeddings, i.e., $v(c) = \vec{c}$.

Averaged word vectors The meaning of some characters may be more clearly expressed within words. From the Wikipedia dataset, we retrieve the k most frequent words $w_1, \ldots, w_k$ that contain the character c. We then average the word vectors of these k words, i.e. defining $v(c) = \frac{1}{k} \sum_{i=1,\ldots,k} \vec{w_i}$.

We assign a score to each candidate character set S by summing the cosine similarity for all its character pairs c_i, c_j:

$$score(S) = \sum_{i,j \leq N, i \neq j} cos(v(c_i), v(c_j)) \quad (1)$$

We then choose the character set that maximizes this score.

4.2 Set-up

We extracted all characters that can be decomposed into two components from the open-source dataset HanziJS.[2] Each character was then assigned to the two character families associated with its two components.[3] We included only characters listed in the *Hanyu Shuiping Kaoshi* (HSK) (Hanban, 2014), the most popular scheme for learning Chinese as a foreign language.

[2]https://github.com/nieldlr/hanzi

[3]Manual inspection would be needed to distinguish between phonetic and semantic components. For a fully automatic algorithm, we relied on the model to learn the distinction rather than manually filtering out phonetic components.

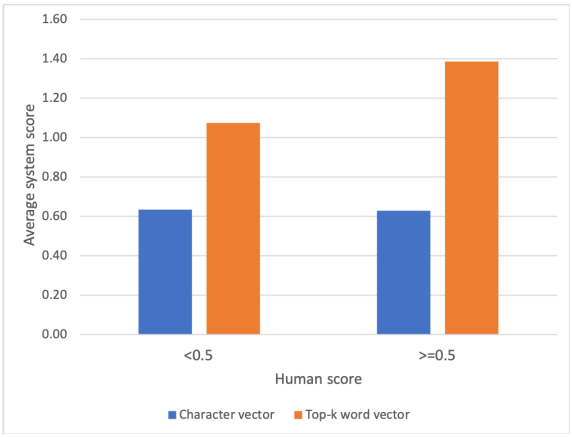

Figure 1: The "Averaged word vectors" method (top-k word vectors) assigns higher scores to similar character pairs (rated 0.5 or over) than dissimilar pairs (rated below 0.5); in contrast, the "Character vector" method does not clearly distinguish the similar and dissimilar pairs.

There were 2,285 characters distributed in 600 character families. We randomly selected ten character families for evaluation. For each family, we constructed two character sets of size $N = 3$ using the two methods for generating $v(c)$ (Section 4.1), with the settings $k = \{5, 10, 15\}$.

4.3 Evaluation

We presented two human judges with all character pairs within the 20 character sets. Both native speakers of Chinese, the judges rated each pair according to the annotation scheme of the SemEval-2012 shared task on Chinese word similarity (Jin and Wu, 2012). The similarity score ranged from 0 (not at all related) to 5 (identical).

We computed the correlation between the averaged human scores and the cosine similarity $cos(v(c_i), v(c_j))$ as generated by the two methods described in Section 4.1. The "Character vector" method attained a Pearson correlation coefficient of only 0.09.[4] The "Averaged word vectors" method achieved a coefficient of 0.80[5], outperforming the "Character vector" method. This coefficient was obtained at $k = 5$, i.e., averaging the 5 most frequent words. Performance degraded at higher values of k, likely because of increased sensitivity to the corpus domain.

To visualize the correlation, we compared the cosine similarity $cos(v(c_i), v(c_j))$ of the similar char-

[4]Correlation with human scores was not significant, at $p > 0.05$

[5]Correlation with human scores was significant, at $p < 0.006$

acter pairs (defined as those with human ratings of 0.5 and above) and the dissimilar pairs. As shown in Figure 1, the "Averaged word vectors" method produces substantially higher similarity scores for the similar pairs than the dissimilar pairs, while the "Character vector" method does not clearly distinguish the similar and dissimilar pairs.

These results suggest that averaged word vectors are more effective for character set construction than direct use of character embeddings. The most frequent words likely play significant roles in shaping the "general" meaning of a character as perceived by native speakers.

5 Subfamilies

We next address Q2 by dividing a family into subfamilies, and evaluating the quality of character sets generated from the subfamilies.

5.1 Approach

We used K-means clustering to produce subfamilies F_i from a character family F. The number of clusters for each family was determined by the silhouette value, which measures the distance between each point in a cluster to the points in its neighboring clusters.[6]

We extracted a character set of size $N = 3$ from each subfamily using the "Averaged word vectors" method at $k = 5$, which obtained the best results (Section 4.3). We identified the two subfamilies F_1 and F_2 with the highest-scoring character sets in terms of $score(S)$, as defined in Section 4.1. For evaluation, we compare the following sets:

Subfamily #1 The character set produced by F_1.

Subfamily #2 The character set produced by F_2.

Mixed Subfamilies The character set produced by randomly swapping one character between Subfamily #1 and Subfamily #2.

Random The character set produced by random selection among characters in F.

5.2 Set-up

Among the 600 character families (Section 4.2), K-means clustering discovered two clusters in 14 character families, and three clusters in 5 character families. Table 1 shows two clusters, or subfamilies, identified in the character family of the component *nǚ* 'female', semantically associated with matrimony and relatives, respectively. We randomly selected eight of these families for evaluation.

Character set	Average score
Subfamily #1	1.67
Subfamily #2	1.32
Mixed Subfamilies	0.75
Random	0.53

Table 2: Human scores on character sets constructed from two subfamilies and two baselines (Section 5.1)

5.3 Evaluation

Similar to the previous experiment, the two human judges rated the similarity of all character pairs in the generated character sets.

As shown in Table 2, the character sets Subfamily #1 (1.67) and Subfamily #2 (1.32) achieved the highest average similarity scores. Both outperformed[7] the Random set, which attained an average of 0.53 only. This result indicates that our proposed method is able to identify characters within a family that are more semantically related than other family members.

Further, Subfamily #1 (1.67) outperformed[8] Mixed Subfamilies (0.75), suggesting that the judges perceived semantic differences between the two subfamilies. Subfamily #2 (1.32) also scored higher than Mixed Subfamilies, although the difference was not significant[9], likely due to the lower degree of similarity between its members compared to their Subfamily #1 counterparts.

6 Conclusion

We have presented the first quantitative study on automatic construction of Chinese character sets to facilitate language learning. We have evaluated a number of methods for character representation and family clustering. Experimental results showed that averaged word vectors achieved statistically significant improvement over direct use of character vectors. Further, K-means clustering produced subfamilies that yielded character sets with distinctive meaning. It is hoped that these methods will help expand the variety and coverage of character

[6] We used the implementation in scikit-learn (Pedregosa et al., 2011). We allowed a maximum of 10 clusters per family, and rejected clusters with less than 5 characters.

[7] Statistically significant at $p < 0.025$ by t-test

[8] Statistically significant at $p < 0.002$ by t-test

[9] At $p = 0.27$

sets for use in CALL applications for Chinese.

Acknowledgements

We gratefully acknowledge support from an Applied Research Grant (project #9667175) at City University of Hong Kong, and a grant from the Hong Kong Institute for Data Science (project #9360163) at City University of Hong Kong.

References

Harald Baayen, Richard Piepenbrock, and Léon Gulikers. 1995. *The CELEX lexical database (CD-ROM)*.

Shaosheng Cao, Wei Lu, Jun Zhou, and Xiaolong Li. 2018. cw2vec: Learning Chinese Word Embeddings with Stroke n-gram Information. In *Proc. 32nd AAAI Conference on Artificial Intelligence*.

Hsueh-Chih Chen, Li-Yun Chang, Kuo-En Chang, Yu-Shiou Chiou, and Yao-Ting Sung. 2011. Chinese Orthography Database and Its Application in Teaching Chinese Characters (in Chinese). *Bulletin of Educational Psychology (Special Issue on Reading)*, 43:269–290.

Flora Hoi Ki Chung and Man Tak Leung. 2008. Data analysis of Chinese characters in primary school corpora of Hong Kong and mainland China: preliminary theoretical interpretations. *Clinical Linguistics and Phonetics*, 22(4-5):379–389.

Mathias Creutz and Krista Lagus. 2006. Morfessor in the Morpho Challenge. In *Proc. PASCAL Challenge Workshop on Unsupervised Segmentation of Words into Morphemes*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. In *Proc. NAACL-HLT*.

Hanban. 2014. *International Curriculum for Chinese Language and Education*. Beijing Language and Culture University Press, Beijing, China.

Peng Jin and Yunfang Wu. 2012. SemEval-2012 Task 4: Evaluating Chinese Word Similarity. In *Proc. First Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 374–377.

H. C. Lam, W. W. Ki, N. Law, A. L. S. Chung, P. Y. Ko, A. H. S. Ho, and S. W. Pun. 2001. Designing CALL for Learning Chinese Characters. *Journal of Computer Assisted Learning*, 17:115–128.

John Lee and Chak Yan Yeung. 2020. Computer-assisted learning for chinese based on character families. In *Information Management and Big Data*, pages 299–305, Cham. Springer International Publishing.

Che Kan Leong, Shek Kam Tse, Ka Yee Loh, and Wing Wah Ki. 2011. Orthographic Knowledge Important in Comprehending Elementary Chinese Text by Users of Alphasyllabaries. *Reading Psychology*, 32(3):237–271.

Y. Li and J. S. Kang. 1993. Analysis of Phonetics of the Ideophonetic Characters in Modern Chinese. In *Information Analysis of Usage of Characters in Modern Chinese (in Chinese)*, pages 84–98, Shanghai. Shanghai Education Publisher.

Susan J. Rickard Liow, Siok Keng Tng, and Cher Leng Lee. 1999. Chinese Characters: Semantic and Phonetic Regularity Norms for China, Singapore, and Taiwan. *Behavior Research Methods, Instruments, and Computers*, 31(1):155–177.

Yanan Lu, Yue Zhang, and Donghong Ji. 2016. Multi-prototype Chinese Character Embedding. In *Proc. 10th International Conference on Language Resources and Evaluation (LREC)*, page 855–859.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

Shek Kam Tse, Ference Marton, Wing Wah Ki, and Elizabeth Ka Yee Loh. 2007. An Integrative Perceptual Approach to Teaching Chinese Characters. *Instructional Science*, 35:375–406.

Jian Xu, Jiawei Liu, Liangang Zhang, Zhengyu Li, and Huanhuan Chen. 2016. Improve Chinese Word Embeddings by Exploiting Internal Structure. In *Proc. NAACL-HLT*, page 1041–1050.

Jinxing Yu, Xun Jian, Hao Xin, and Yangqiu Song. 2017. Joint Embeddings of Chinese Words, Characters, and Fine-grained Subcharacter Components. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 286–291.