ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20210128043310
Original scientific papers

A Machine Learning Classification Algorithm for Vocabulary Grading in Chinese Language Teaching

Yinbing ZHANG, Jihua SONG*, Weiming PENG*, Dongdong GUO, Tianbao SONG

Abstract: Vocabulary grading is of great importance in Chinese vocabulary teaching. This paper starts with an analysis of the lexical attributes that affect lexical complexity, followed by an explanation of the extraction of lexical attribute information combined with the constructed word-formation knowledge base, the construction of mapping functions corresponding to lexical attributes, and the quantitative representation of the attributes that form the basis for vocabulary grading. Based on this, a machine learning classification algorithm is creatively applied to the Chinese vocabulary grading problem. Using the comparative analysis of vocabulary grading models based on common machine learning classification algorithms, the importance measurement analysis of Chinese vocabulary attributes based on different feature selection methods is performed, and a vocabulary grading model is constructed based on the machine learning classification algorithm and feature importance selection of different feature selection algorithms. A comparison of the experimental results demonstrated that the classification model based on the support vector machine (SVM) algorithm and top six attribute groups by the importance of feature selection received the best effect. To improve vocabulary grading, a variety of feature selection algorithms were used to fuse the importance of lexical attributes on average. Then an experiment was conducted for vocabulary grading combined with the Bagging + SVM integration algorithm and top six attribute groups by the importance of feature selection. The experimental results demonstrated that the combination scheme achieved a better effect.

Keywords: integration algorithm; machine learning classification algorithm; vocabulary grading; word-formation knowledge base

1 INTRODUCTION

In essence, language is a tool for human communication. Whether people can communicate depends on the understanding of semantics, and words are the most important carrier of semantics and the only way to learn a language well. As Chomsky [1] indicated in his discussion on language systems, a person who has a language has access to detailed information about the words of the language. Vocabulary is one of the three elements of language, and the important content for learning and mastering a language. The importance of vocabulary learning is self-evident. David Wilkins [2] said that "without grammar very little can be conveyed, without vocabulary nothing can be conveyed".

In the international Chinese language teaching field, the new HSK syllabus is the main basis and guiding document for the proposition and test of Chinese-language ability. It has a wide influence, and after several revised editions, it has been greatly improved compared with the original edition. However, it still has some shortcomings. As Zhang Jinjun [3] indicated, the design of the vocabulary level still needs to be improved, including which words should be accepted, which words should be abandoned, and which words should be placed at a higher level, all of which need further detailed research and adjustment. Each revision and adjustment of the syllabus consumes a great deal of manpower and material resources. Despite this, for various reasons, each revision may not achieve satisfactory results. Therefore, it is necessary and important significant to study vocabulary grading in the Chinese language teaching field.

The vocabulary grading problem based on different lexical attributes can be regarded as a classification problem; that is, each word is mapped to the attribute vector of its corresponding attribute element by the extraction of the characteristic attribute value of the word, then the classification algorithm model in machine learning is used to determine the level of the word, and the word is divided into six levels from level 1 to level 6. In this study, we use a common machine learning classification algorithm based on the lexical attribute vector set to conduct a vocabulary grading experiment.

During the experiment, combined with the features widely used in [4-8] to assess the comprehensive complexity of vocabulary, the cross-validation method is used to predict the vocabulary levels of Chinese language teaching; that is, the vocabulary and its corresponding attribute vector set are divided into two parts; one part is used as the training set to train the classifier, and the other is used as the validation set to test the performance of the classification model. The schematic diagram of the vocabulary grading model based on the machine learning classification algorithm used in this study is shown in Fig. 1.

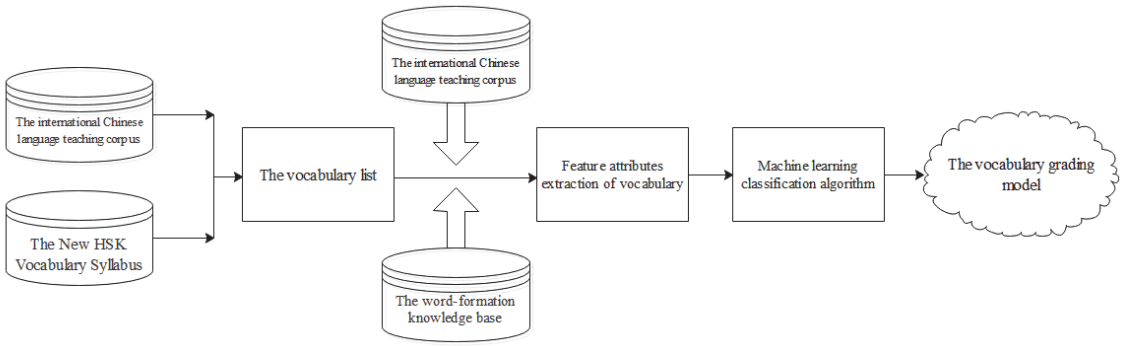


Figure 1 Machine Learning Classification Model for Vocabulary Grading in Chinese Language Teaching

2 RELATED WORK

The machine learning classification algorithm is widely used in the field of natural language processing. There have been many studies on, for example, text classification, text readability prediction, complex vocabulary recognition, and vocabulary grading.

Regarding text classification, Pingpeng Yuan et al. [9] proposed a multi-class text classification algorithm called SVM-KNN by combining a support vector machine (SVM) with the KNN classification algorithm. First, SVM is used to recognize the category boundary, then KNN is used to classify the document boundary, which overcomes the defects of SVM and KNN, and improves the performance of multi-category text classification. In [10], a new text classification algorithm based on thesaurus knowledge and the KNN classification algorithm was proposed, and good classification results were obtained. In [11], a text classification algorithm was proposed based on a backpropagation neural network (BPNN) and modified BPNN (MBPNN). In [12], the text classification algorithms based on machine learning were summarized, and the principles and applications of text classification, text clustering, and text mining were introduced in detail.

Regarding text readability research, Kate Rohit [13] discussed the application of common machine learning classification algorithms in text readability prediction, and the natural language text readability prediction system constructed received relatively satisfactory results. Sarah Schwarm et al. [14] combined SVM with traditional reading-level measurement, a statistical language model, and other language processing tools to predict the readability of English news text, and achieved better prediction results than traditional methods. Sun Gang [15] proposed a text readability prediction method based on linear regression after comprehensively considering the surface features, lexical features, grammatical features, and other text features, and demonstrated the effectiveness of the method through experiments.

Regarding complex vocabulary recognition, Matthew Shardlow [16] compared the automatic recognition of complex vocabulary using different methods. Compared with other methods, the accuracy of the automatic recognition of complex vocabulary based on SVM slightly improved. Lucia Special et al. [17] trained an SVM, which can sort words according to their complexity to select the words that need to be simplified. Muralidhar Pantula et al. [18] trained a machine learning classification model with 19 internal vocabulary features, and achieved 84.75% accuracy on the given experimental dataset.

Regarding vocabulary syllabus development or vocabulary grading, Gala et al. [19] selected 9 of the 27 internal attributes of vocabulary that can best predict the vocabulary level, trained an SVM classifier, and used the method of five-fold cross-validation on the experimental data. The average accuracy of the three classification results was 62%. Additionally, Gala et al. [20] conducted a comparative study of the classification of words in MANULEX and FLELEX based on an SVM. They used four categories of 49 features, including spelling features, morphological features, semantic features, and statistical features, to train three and six categories of SVM classifiers. The accuracies of the final experimental results were 63% and 43%, respectively. The study in [21] showed that the length of English words does not seem to predict lexical complexity, and only two frequency features were used to obtain very good results.

The purpose of this study is to explore the application of a machine learning classification algorithm in Chinese teaching vocabulary grading based on the analysis of lexical attributes. In Section 3, the data resource for this study is introduced, the extraction of lexical attribute information combined with the constructed word-formation knowledge base is completed, the mapping functions corresponding to the lexical attributes are constructed, and the quantitative representation of the attributes that form the basis for vocabulary grading is obtained. In Section 4, first, the evaluation index of the effect of the vocabulary classification model is provided. Then, a comparative analysis of the experimental results of vocabulary grading based on different classification algorithms is performed. Additionally, to obtain a better effect, the average fusion of the importance of lexical attributes is conducted for a variety of feature selection algorithms. Finally, a summary of this study is provided in Section 5.

3 EXPERIMENTAL DATA ACQUISITION

To convenient for the comparison and analysis of the experimental results, we need to choose a comparison object that has high recognition or authority. For this study, we choose the intersection of vocabulary covered by the eight sets of textbooks planned by Hanban and the vocabulary of the new HSK syllabus as the research object, which includes 2885 words. Detailed information about the eight selected textbooks is shown in Tab. 1.

To study vocabulary grading, it is necessary to analyze the lexical attributes that affect it. Based on previous studies, we choose the Chinese character word formation attribute, general vocabulary attributes, and statistical vocabulary attribute as the characteristic attributes of Chinese vocabulary's comprehensive complexity.

Table 1 Basic information of 8 sets planning textbooks

Book Name	Copy Number	Article Number	Sentence Number	Word Number
Great Wall Chinese(Essentials in Communication)	6	180	2138	20610
Mandarin Teaching Toolbox	3	22	92	746
Chinese Paradise	6	50	174	1089
Happy Chinese	3	94	484	6329
Contemporary Chinese	4	86	1716	26367
Learn Chinese with Me	4	176	1687	21820
New Practical Chinese Reader	6	139	4175	77758
Integrated Chinese	4	60	1585	24539
Total	36	807	12051	179258

The Chinese character word formation attribute includes average strokes and structural types of Chinese characters; general vocabulary attributes include part of speech (POS), number of syllables, and word-formation structure; statistical vocabulary attributes include frequency, number of word senses, average number of morpheme senses, and morpheme word formation ability.

To apply lexical attributes to the quantitative representation of comprehensive vocabulary complexity, we need to map the lexical attributes using the mapping function. To eliminate the difference in dimensions and the value range between different attribute representations, and make the mapping value of the mapping function meet the requirements of standardization, we set the mapping value between 0 and 1 while constructing the mapping function.

3.1 Word Formation Chinese Character Attribute

(1) Average strokes. Average strokes refers to the average number of strokes of all Chinese characters in a word. For example, the average strokes of "中(zhōng; middle)" is 4, the average strokes of "什么(shénme; what)" is 3.5, and the average strokes of "消耗(xiāohào; consume)" is 10.

In English, the number of letters in a word is often used to evaluate the complexity of a word, which is discussed in [22]. Similarly, the average strokes of a word in Chinese has an important influence on the complexity of the word. Combined with the average strokes in words in the corpus and the proportion of each vocabulary level in the HSK syllabus, the average strokes' piecewise mapping function $f_1(x)$ is defined, as shown in Tab. 2.

Table 2 Average strokes piecewise mapping function

x	[1,4)	4	(4,5]	(5,6]	(6,8]	(8,23]
$f_1(x)$	1/6	1/3	1/2	2/3	5/6	1

(2) Structural types of Chinese characters. In this study, structural types of Chinese characters refers to the first structure of Chinese characters, which has 13 structure types: left right structure, upper and lower structure, left middle right structure, upper middle lower structure, whole surround structure, upper left surround structure, lower left surround structure, upper right surround structure, lower opening frame surround structure, right opening frame surround structure, upper opening frame surround structure, overlapping structure, and single structure.

In [23], a detailed statistical analysis was performed of 2905 Chinese characters' structure types in the list of "graded Chinese characters". The corresponding relationship between structural types of Chinese characters, storage symbols, and examples in this study is shown in Tab. 3.

Considering the frequency distribution and the proportion of each vocabulary level in the HSK syllabus, the structural types of Chinese characters' piecewise mapping function $f_2(x)$ is defined, as shown in Tab. 4.

Table 3 Chinese characters structural types storage symbols and examples

No.	Storage symbols	Structural types	Examples
1	⿰	Left right structure	的(de; of), 你(nǐ; you), 好(hǎo; good)
2	⿱	Upper and lower structure	多(duō; many), 想(xiǎng; think), 要(yào; want)
3	⿲	Left middle right structure	班(bān; class), 辩(biàn; debate), 鞭(biān; whip)
4	⿳	Upper middle lower structure	尝(cháng; taste), 宽(kuān; wide), 害(hài; harm)
5	⿴	Whole surround structure	国(guó; country), 回(huí; return), 图(tú; picture)
6	⿸	Upper left surround structure	在(zài; stay), 看(kàn; look), 房(fang; room)
7	⿺	Lower left surround structure	还(hái; still), 这(zhè; this), 过(guò; pass)
8	⿹	Upper right surround structure	可(kě; can), 司(sī; department), 句(jù; sentence)
9	⿷	Lower opening frame surround structure	问(wèn; ask), 间(jiān; between), 同(tóng; with)
10	⿵	Right opening frame surround structure	医(yī; doctor), 区(qū; area), 巨(jù; huge)
11	⿶	Upper opening frame surround structure	凶(xiōng; fierce), 山(shān; mountain), 义(yì; justice)
12	⿻	Overlapping structure	来(lái; come), 夹(jiá; clip), 爽(shuǎng; clear)
13	Space	Single structure	我(wǒ; I), 了(le; has been), 不(bú;no), 也(yě; also)

Table 4 Chinese characters structural types piecewise mapping function

x	⿰	⿱	other
$f_2(x)$	1/3	2/3	1

The structural types of a Chinese character's mapping value for a word are represented by the average value of each Chinese character in the word. For example, the structural types of the Chinese character sequence of "现在(xiànzài;now)" is "⿰⿸" and its mapping value is (1/3 + 1)/2 = 0.6667.

3.2 Vocabulary General Attributes

(1) POS. Generally, Chinese POS are divided into 14 categories: nouns, verbs, adjectives, numerals, quantifiers, pronouns, distinguishing words, adverbs, prepositions, conjunctions, auxiliaries, interjections, modal words, and onomatopoeia. The POS and tagging symbol set used in this study are shown in Tab. 5.

Table 5 POS and tagging symbol set

Id	POS	Tag	Id	POS	Tag
1	Nouns	***n***	9	Adverbs	***d***
2	time words	***t***	10	Prepositions	***p***
3	Localizers	***f***	11	Conjunctions	***c***
4	Numerals	***m***	12	Auxiliary	***u***
5	Quantifiers	***q***	13	Interjections	***e***
6	Pronouns	***r***	14	Onomatopoeia	***o***
7	Verbs	***v***	15	Default	***x***
8	Adjectives	***a***			

Considering the proportion distribution of each POS, the POS piecewise mapping function $f_3(x)$ is defined as shown in Tab. 6.

Table 6 POS piecewise mapping function

x	v, n	a, d	other
$f_3(x)$	1/3	2/3	1

(2) Number of syllables. In this study, the number of syllables refers to the number of Chinese characters contained in a word. For Chinese vocabulary, disyllabic vocabulary is the most common, accounting for the largest proportion, and monosyllabic vocabulary is the most complex.

Table 7 Syllable number piecewise mapping function

x	2	3, 4	1
$f_4(x)$	1/3	2/3	1

According to the statistical distribution of the number of syllables, the number of syllables' piecewise mapping function $f_4(x)$ is defined as shown in Tab. 7.

(3) Word-formation structure. The word-formation structure of a word refers to the relationship between morphemes, for example, coordinate, attributive-centered, endocentric adverbial, predicate-complement, verb-object, subject-predicate, and overlapping. In previous studies, a symbol set designed by Guo et al. [24-25] was used to describe the word-formation structure, as shown in Tab. 8, and we completed 63,193 record scale word-formation structure annotations based on the Modern Chinese Dictionary [26].

Table 8 Word-formation structure symbol set

Structural relationship	Tag	Example
Coordinate	…	花…草(huācǎo; flowers and plants); 风…雨(fēngyǔ; wind and rain); 父…母(fùmǔ; parent)
Attributive-centered	↗	鸡↗蛋(jīdàn;egg); 米↗饭(mǐfàn;rice); 中↗文(zhōngwén;Chinese)；彭↗总(péngzǒng; Mr.Peng)
Endocentric adverbial	→	极→具(jíjù; Extremely); 深→感(shēn gǎn; depth perception); 好→看(hǎokàn; nice)
Predicate-complement	←	赶←跑(gǎn; drive away); 看←清(kànqīng; see clearly); 拿←下(náxià; lift down)
Verb-object	|	赚|钱(zhuànqián; make money); 做|饭(zuòfàn; cook); 征|税(zhēngshuì; taxation)
Subject-Predicate	‖	兵‖变(bīngbiàn; mutiny); 唇‖裂(chúnliè;cleft lip); 胆‖怯(dǎnqiè; timidity); 地‖震(dìzhèn; earthquake)
Overlapping	·	谢·谢(xièxiè; thank you); 妈·妈(māmā; mom); 往·往(wǎngwǎng; often); 人·人(rénrén; everyone)
Other	-	桌-上(zhuōshàng; on the table); 两-只(liǎngzhī; two); 一-整-套(yīzhěngtào; a full set of) 看-了(kànle;looked); 看-着(kànzhe;look); 看-过(kànguò; haveseen) 拿-得-起(nádéqǐ; be able to take up); 华山-之-巅(huáshānzhīdiān; top of Huashan Mountain)

According to the statistical distribution of the vocabulary word-formation structure and the proportion of each vocabulary level in the HSK syllabus, the word-formation structure's piecewise mapping function $f_5(x)$ is constructed as shown in Tab. 9, and if the annotation is empty, this means that no word-formation structure is annotated.

Table 9 Word-formation structure piecewise mapping function

x	empty	…, ↗	other
$f_5(x)$	1/3	2/3	1

3.3 Vocabulary Statistical Attributes

(1) Frequency. Frequency refers to the frequency of words that appear in a specific corpus. In many studies, the results have shown that there is a close relationship between the frequency of words and their complexity, which is perhaps the most commonly used vocabulary attribute to express complexity [27]. In this study, word frequency statistics are based on the above listed eight sets of textbooks. According to the proportion of each vocabulary level in the HSK syllabus and the word frequency in the corpus, the frequency attribute's piecewise mapping function $f_6(x)$ is constructed as shown in Tab. 10.

Table 10 Frequency attribute piecewise mapping function

x	[190,4719]	[90,190)	[40,90)	[17,40)	[5,17)	[1,5)
$f_6(x)$	1/6	1/3	1/2	2/3	5/6	1

(2) Number of word senses. In this study, the number of word senses refers to the number of different senses of words in the Modern Chinese Dictionary. The lower the number of word senses, the easier it is to understand the word and the lower the difficulty; by contrast, the greater the number of word senses, the more difficult it is to distinguish and understand the word. Based on this idea, combined with the distribution of the number of word senses and the proportion of each vocabulary level in the HSK syllabus, the mapping function $f_7(x)$ of the number of word senses is defined, as shown in Tab. 11.

Table 11 Number of word senses attribute piecewise mapping function

x	1	2	3	4	5,6	(6,27]
$f_7(x)$	1/6	1/3	1/2	2/3	5/6	1

(3) Average number of morpheme senses. The average number of morpheme senses refers to the average number of each morpheme sense of a word. Through statistical analysis, the mapping function $f_8(x)$ of the average number of morpheme senses is defined, as shown in Tab. 12.

Table 12 Average number of word senses of morphemes attribute piecewise mapping function

x	[0.5, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 7)	[7, 27]
$f_8(x)$	1/6	1/3	1/2	2/3	5/6	1

(4) Morpheme word formation ability. The morpheme word formation ability refers to the average number of times that morphemes of words appear as morphemes of other words. Based on the statistical distribution of the average morpheme word formation ability and the proportion of each vocabulary level in the HSK syllabus, the mapping function $f_9(x)$ of the morpheme word formation ability is defined, as shown in Tab. 13.

Table 13 Morpheme word formation ability attribute piecewise mapping function

x	[430, 1116]	[311, 430)	[227, 311)	[160, 227)	[90, 160)	[1,90)
$f_9(x)$	1/6	1/3	1/2	2/3	5/6	1

The vocabulary attribute values of some words in this study are shown in Tab. 14, and the corresponding mapping values are shown in Tab. 15.

Table 14 Vocabulary attribute values of some words

id	ciyu	bh_avg	jg_seq	pos	yj_sum	gc_gx	fre	ciyx_sum	ysyx_avg	ysgc_avg
1	时候(shí hòu; time)	8.5	□□	名	2	…	304	2	9	126.5
2	怎么(zěn me; how)	6	日日	代	2	-	292	4	3	15
3	知道(zhī dào; know)	10	□口	动	2	|	200	1	12	198
4	朋友(péng yǒu; friend)	6	□口	名	2	…	194	2	4	33.5
5	问题(wèn tí; problem)	10.5	门口	名	2	↗	159	5	4.5	81
6	起来(qǐ lái; get up)	8.5	口□	动	2	←	102	7	18	152
7	提高(tí gāo; raise)	11	□□	动	2	←	23	1	9.5	164
8	希望(xī wàng; hope)	9	日日	动	2	…	73	3	8	56.5
9	博物馆(bó wù guǎn; museum)	10.33	□□□	名	3	↗	19	1	3	118
10	售货员(shòu huò yuan; salesperson)	8.667	日日日	名	3	|↗	2	1	4	80. 7

Table 15 Vocabulary attribute mapping values of some words

id	ciyu	bh_avg	jg_seq	pos	yj_sum	gc_gx	fre	ciyx_sum	ysyx_avg	ysgc_avg
1	时候(shí hòu; time)	1.0000	0.6667	0.3333	0.3333	0.6667	0.1667	0.3333	1.0000	0.8333
2	怎么(zěn me; how)	0.6667	0.6667	1.0000	0.3333	1.0000	0.1667	0.6667	0.5000	1.0000
3	知道(zhī dào; know)	1.0000	0.6667	0.3333	0.3333	1.0000	0.1667	0.1667	1.0000	0.6667
4	朋友(péng yǒu; friend)	0.6667	0.6667	0.3333	0.3333	0.6667	0.1667	0.3333	0.6667	1.0000
5	问题(wèn tí; problem)	1.0000	1.0000	0.3333	0.3333	0.6667	0.3333	0.8333	0.8333	1.0000
6	起来(qǐ lái; get up)	1.0000	1.0000	0.3333	0.3333	1.0000	0.3333	1.0000	1.0000	0.8333
7	提高(tí gāo; raise)	1.0000	0.6667	0.3333	0.3333	1.0000	0.6667	0.1667	1.0000	0.6667
8	希望(xī wàng; hope)	1.0000	0.6667	0.3333	0.3333	0.6667	0.5000	0.5000	1.0000	1.0000
9	博物馆 (bó wù guǎn; museum)	1.0000	0.3333	0.3333	0.6667	0.6667	0.6667	0.1667	0.5000	0.8333
10	售货员 (shòu huò yuan; salesperson)	1.0000	0.6667	0.3333	0.6667	1.0000	1.0000	0.1667	0.6667	1.0000

4 EXPERIMENT AND ANALYSIS

4.1 Evaluation Index

Combined with the purpose of vocabulary grading, this research selects several evaluation indexes to compare and evaluate the prediction results of the classification algorithm, such as accuracy, approximate accuracy, root mean square error, Kappa coefficient, Pearson correlation coefficient and so on.

(1) Accuracy. Accuracy of vocabulary grading refers to the proportion of the number of words whose grades are consistent with those in the vocabulary syllabus to the total number of predicted words. The calculation formula of accuracy for n samples is as follows:

$$accuracy(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} \text{level}(\hat{y}_i = y_i) \tag{1}$$

where y is the actual level of words in the syllabus, $\hat{y}$ is the prediction level of words, and n_{samples} is the number of words involved in the evaluation.

(2) Approximate accuracy. Approximate accuracy of vocabulary grading refers to the proportion of the number of words whose grades are similar to those in the syllabus to the total number of predicted words. The grades are similar, here it is defined that the absolute value of the difference between the two is not more than 1. The calculation formula of approximate accuracy is as follows:

$$appaccuracy(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} \text{level}(|\hat{y}_i - y_i| \le 1) \tag{2}$$

(3) Root mean square error (*RMSE*). Root mean square error is the square root of the average square error between the predicted and the actual vocabulary level, which is used to measure the deviation between the predicted and the actual vocabulary level. The calculation formula of root mean square error is as follows:

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n_{\text{samples}}} \sum_{i=1}^{n_{\text{samples}}} (\hat{y}_i - y_i)^2} \tag{3}$$

(4) Kappa coefficient (*KC*). Kappa coefficient is a method used to evaluate the consistency in statistics. It can be used to evaluate the accuracy of multi classification models. It represents the proportion of errors in classification results that are less than those in completely random classification results. The calculation formula of Kappa coefficient is as follows:

$$K = \frac{n_{\text{samples}} \sum_{i=1}^{C} x_{i,i} - \sum_{i=1}^{C} (a_i \cdot b_i)}{n_{\text{samples}}^2 - \sum_{i=1}^{C} (a_i \cdot b_i)} \tag{4}$$

where n_{samples} is the number of words involved in the experiment, $x_{i,i}$ is the number of words correctly predicted in class i, which corresponding to the elements on the diagonal of the prediction confusion matrix, and C is the number of vocabulary levels, a_i is the actual number of words at level i, b_i is the predicted number of words at level i.

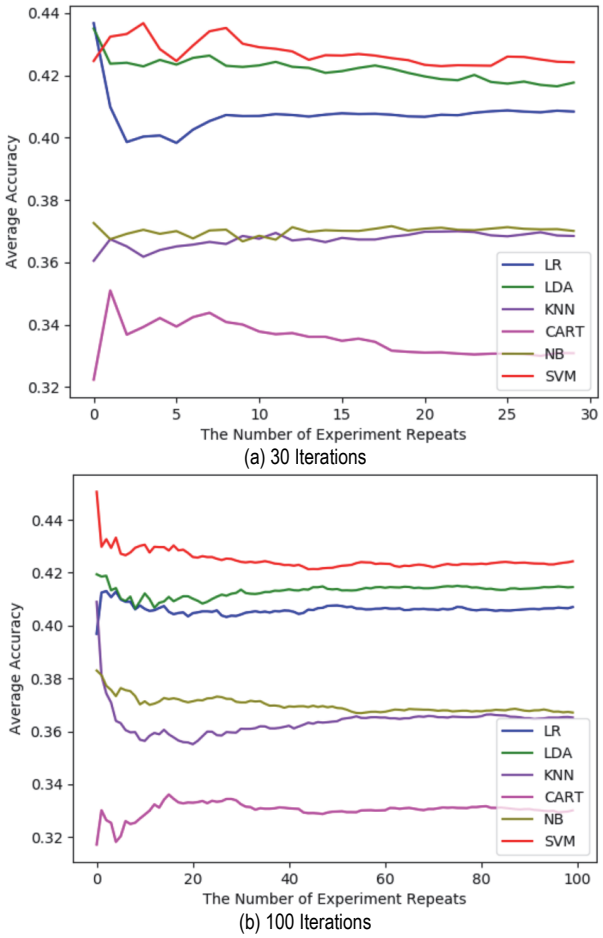
(5) Pearson correlation coefficient *(PCC*). Pearson correlation coefficient is the most commonly used correlation coefficient in statistics, which is used to measure the linear correlation between two variables. The calculation formula of Pearson correlation coefficient is as follows:

$$\rho(y,\hat{y})=\frac{\sum_{i=1}^{n_{\text{samples}}}(\hat{y}_i-\bar{\hat{y}})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n_{\text{samples}}}(\hat{y}_i-\bar{\hat{y}})^2}\sqrt{\sum_{i=1}^{n_{\text{samples}}}(y_i-\bar{y})^2}} \quad (5)$$

where y is the actual level of words in the syllabus, and $\bar{y}$ is the average of y. $\hat{y}$ is the prediction level of words, and $\bar{\hat{y}}$ is the average of $\hat{y}$. n_{samples} is the number of words involved in the evaluation.

4.2 Comparison and Analysis of the Experimental Results Based on Several Common Machine Learning Classification Algorithms

Several commonly used machine learning classification algorithms were used in the experiment: *LR*, *LDA*, *KNN*, *CART*, *NB*, and *SVM* [28]. The experimental data were randomly divided, with 80% of data in the training set and 20% in the test set. Repeated experiments were conducted, and the average value of the experimental results was used to evaluate the classification model. The following is a comparative analysis of the classification results based on the new HSK syllabus vocabulary level.

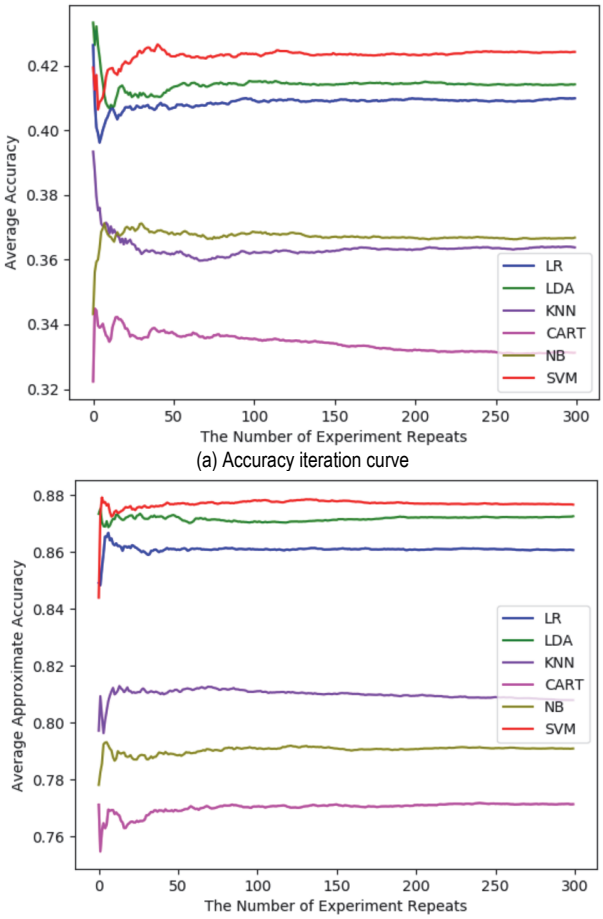


(a) 30 Iterations

(b) 100 Iterations

Figure 2 Iterative Curve of Average Accuracy

To observe the stability of the classifier, based on various classification algorithms, the average accuracy and average approximate accuracy were compared through 30, 100, and 300 iterations. Based on the new HSK syllabus vocabulary level, the average accuracy of 30 iterations and 100 iterations is shown in Fig. 2, the average accuracy and average approximate accuracy of 300 iterations is shown in Fig. 3, and detailed information based on each evaluation index of 300 iterations is shown in Tab. 16.



(a) Accuracy iteration curve

(b) Approximate accuracy iteration curve

Figure 3 Iterative Curve of 300 Iterations Experiments

Fig. 3a shows that the accuracy of the six classification algorithms based on the new HSK syllabus vocabulary level from high to low was *SVM*, *LDA*, *LR*, *KNN*, *NB*, and *CART*, and similar sorting results were achieved for the approximate accuracy in Fig. 3b.

Table 16 Experimental results of vocabulary classification based on the new HSK syllabus vocabulary level

Classification algorithm	Accuracy	Approximate accuracy	*RMSE*	*KC*	*PCC*
LR	40.99%	85.92%	1.1320	0.1939	0.6449
LDA	41.42%	87.24%	1.0867	0.2074	0.6751
KNN	36.38%	80.74%	1.2584	0.1733	0.5868
CART	33.13%	77.21%	1.3375	0.1954	0.5552
NB	36.68%	78.96%	1.2932	0.1953	0.6419
SVM	42.42%	87.61%	1.0623	0.2307	0.6764

The experimental results in Tab. 16 show that *SVM* achieved the best effect of the six classification algorithms in terms of accuracy, which was 42.42%, and approximate accuracy, which was 87.61%. *RMSE* was 1.0623, *KC* was 0.2307, and *PCC* was 0.6764. However, generally, the accuracy of the six classification algorithms was not very high. The reason is that the vocabulary of the HSK syllabus was divided into six levels from level 1 to level 6, and the

level division was relatively fine. It was reasonable to divide a specific word into the current level or adjacent level, and this was also verified by another analysis index: "approximate accuracy". The approximate accuracies of the six classification models were greater than 75%: *SVM* and *LDA* were 87.61% and 87.24%, respectively.

Compared with the results in [30], the accuracy and approximate accuracy of the six classification algorithms in this study were all greater, to a certain extent. Particularly, for the *SVM* classification algorithm, the accuracy was 42.42%, which is 14.38% higher than the accuracy of 28.04% in [30]; and the approximate accuracy was 87.61%, which is 15.13% higher than the approximate accuracy of 72.48% in [30].

4.3 Experimental Effect Improvement

Each classification algorithm has its own characteristics. To improve the experimental results, we can improve two aspects of the experiment: feature selection of lexical attributes and classification algorithm integration.

4.3.1 Classification Effect Improvement Based on Feature Importance Selection of Different Feature Selection Algorithms

To make full use of fewer features and improve the classification effect, it is necessary to select features. Through feature selection, we can reduce the number of features, and reduce the influence of irrelevant features and redundant features on the classification effect, which is more conducive to the understanding of the influencing factors of classification. Simultaneously, we can reduce the space and time costs, and improve performance. In [29], the authors indicated that the appropriate feature selection algorithm and the appropriate number of feature subsets do not affect the classification effect of the classifier, or even improve the classification effect.

Feature selection methods can be divided into three categories: filter, wrapper, and embedding. For the calculation of the importance of lexical attributes, different feature selection algorithms obtain different importance measurements. In this study, a variety of feature selection methods were used to calculate the importance of lexical attributes. To eliminate the influence of different feature selection algorithms and improve the rationality of the importance measurement of lexical attributes, it was necessary to standardize the importance calculated by each feature selection algorithm. The min-max standardization method was used for standardization; the calculation formula is as follows:

$$y_i = \frac{x_i - \min\limits_{1\le j\le n}\{x_j\}}{\max\limits_{1\le j\le n}\{x_j\} - \min\limits_{1\le j\le n}\{x_j\}} \tag{6}$$

Based on the results calculated, the average value was used to express the importance of the corresponding lexical attributes. The details are shown in Tab. 17.

Table 17 Importance measurement of lexical attributes based on different feature selection methods

No.	Lexical attributes	Filter			Wrapper	Embedding		Average
		X^2-test	*F*-test	MI	RFE	ET	RF	
1	Average strokes	0.0828	0.0401	0.0528	0.7500	0.6484	0.6304	0.3674
2	Chinese characters structural types	0.0454	0.0130	0.0233	0.8750	0.5678	0.6699	0.3657
3	POS	0.0621	0.0117	0.0243	0.3750	0.2703	0.3343	0.1796
4	Syllable number	0.3500	0.0519	0.0757	0.5000	0.0000	0.0000	0.1629
5	Word-formation structure	0.1071	0.0249	0.0572	0.2500	0.1823	0.2071	0.1381
6	Frequency	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
7	Number of word senses	0.5283	0.0669	0.0905	0.6250	0.4020	0.4548	0.3613
8	Average number of morpheme senses	0.0000	0.0000	0.0000	0.0000	0.3944	0.4381	0.1388
9	Morpheme word formation ability	0.0182	0.0096	0.0192	0.1250	0.5557	0.5903	0.2197

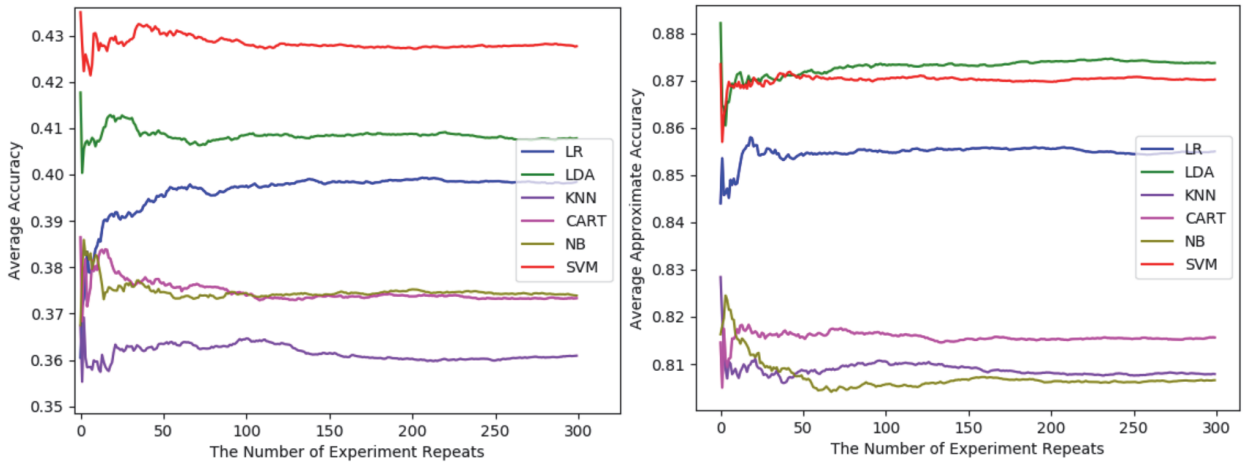
Tab. 17 shows that, according to the measurement results of feature importance for different feature selection methods, the order of importance of the number of lexical attributes was 6, 1, 2, 7, 9, 3, 4, 8, and 5. Through comparative analysis of the experiments, the best classification effect was obtained when the top six lexical attributes of importance were selected. According to the importance of the lexical attributes obtained from expert knowledge in [30], the top six numbers of lexical attributes were 6, 2, 7, 4, 8, and 1. A comparison of the two ranking results of lexical attribute importance demonstrates that there are some differences in the order of lexical attribute importance obtained by the two approaches. The comparison results for the three lexical attribute groups are shown in Tab. 18.

As shown in Tab. 18, a comparison of the classification results of the three attribute groups demonstrates that, considering all attributes, the LR and LDA classification algorithm achieved relatively good classification results based on the top six attribute groups in [30]; KNN and the CART classification algorithm achieved relatively good classification results; and based on feature selection for the top six attribute groups, NB and the SVM classification algorithm achieved relatively good classification results. Among all the experimental results, the SVM classification algorithm based on feature selection for the top six attribute groups achieved the best classification effect, and the accuracy reached 42.76% and the approximate accuracy reached 87.02%. Compared with the accuracy of 28.04% and approximate accuracy of 74.48% in [30], the classification effect greatly improved. A comparison of the experimental results demonstrated that the classification model based on the SVM algorithm and top six attribute groups by the importance of feature selection received the best effect. The reasons can be summarized into two aspects. Through feature selection, we could choose the set

of lexical attributes that contributed more to classification from the combination of lexical attributes, which was more conducive to the improvement of the classification effect. By contrast, the SVM had good learning ability for small sample and high-dimension classification, and obtained a low error rate and made good classification decisions for data points outside the training set. For the six classification algorithms, the iterative curve for accuracy and approximate accuracy based on feature selection for the top six attribute groups is shown in Fig. 4.

Table 18 Comparison of experimental results based on feature selection

Classification algorithm	Combination of lexical attributes	Accuracy	Approximate accuracy	*RMSE*	*KC*	*PCC*
LR	All Attributes	40.99%	85.92%	1.1320	0.1939	0.6449
	Top 6 attribute groups in reference [30]	39.72%	86.02%	1.1359	0.1688	0.6399
	Top 6 attribute groups based on FS	39.84%	85.50%	1.1361	0.1803	0.6488
LDA	All Attributes	41.42%	87.24%	1.0867	0.2074	0.6751
	Top 6 attribute groups in reference [30]	40.62%	87.49%	1.0763	0.1955	0.6816
	Top 6 attribute groups based on FS	40.78%	87.37%	1.0854	0.1995	0.6776
KNN	All Attributes	36.38%	80.74%	1.2584	0.1733	0.5868
	Top 6 attribute groups in reference [30]	36.76%	81.36%	1.2230	0.1829	0.6129
	Top 6 attribute groups based on FS	36.09%	80.79%	1.2285	0.1775	0.6128
CART	All Attributes	33.13%	77.21%	1.3375	0.1954	0.5552
	Top 6 attribute groups in reference [30]	39.04%	83.54%	1.1772	0.2051	0.6393
	Top 6 attribute groups based on FS	37.33%	81.546%	12357	0.1896	0.6173
NB	All Attributes	36.68%	78.96%	1.2932	0.1953	0.6419
	Top 6 attribute groups in reference [30]	36.72%	79.55%	1.2706	0.1947	0.6625
	Top 6 attribute groups based on FS	37.39%	80.66%	1.2485	0.1983	0.6688
SVM	All Attributes	42.42%	87.61%	1.0623	0.2307	0.6764
	Top 6 attribute groups in reference [30]	41.66%	88.13%	1.0541	0.2198	0.6764
	Top 6 attribute groups based on FS	42.76%	87.02%	1.0628	0.2451	0.6835



(a) Accuracy Iteration Curve (b) Approximate Accuracy Iteration Curve

Figure 4 Iterative Curve of Experimental Results Based on Feature Selection Top 6 Attribute Groups

4.3.2 Vocabulary Grading Effect Improvement Based on the Integration Algorithm

Based on the classification algorithm used in the above experiment, combined with the Bagging integration algorithm, the effect of vocabulary grading improved. The integration algorithm is a common method used to improve the experimental effect. At present, popular integration algorithms mainly include the bagging algorithm, boosting algorithm, and voting algorithm.

(1) Vocabulary grading effect improvement based on CART and the bagging integration algorithm

To measure the importance of lexical attributes based on CART, the "DecisionTreeClassifier()" model in "sklearn.tree" of Python was selected, and the "feature_importances_" property of the model was used to represent the importance of lexical attributes. Because the results returned each time were slightly different, to express the importance of lexical attributes more stably, the average value was calculated by repeating the experiment 300 times, and the calculation results were standardized using the min-max standardization method. The importance measure results for the lexical attributes are shown in Tab. 19.

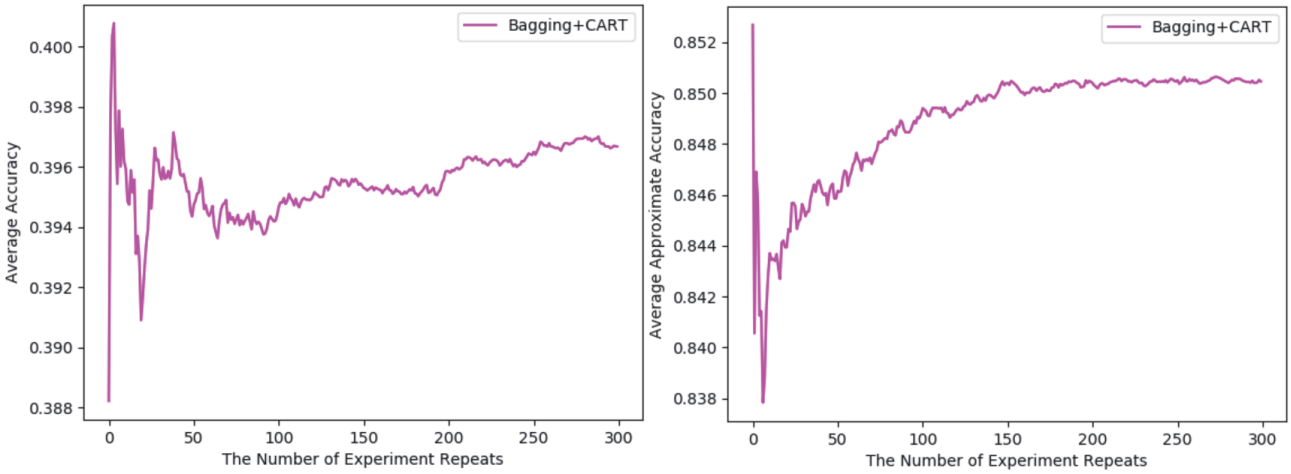
Table 19 Importance measurement of lexical attribute features based CART

Lexical attributes No.	1	2	3	4	5	6	7	8	9
Importance measurement	0.6598	0.6323	0.3008	0.0000	0.3684	1.0000	0.6119	0.4461	0.6154

The bagging algorithm used in the experiment was implemented using the "Bagging Classifier" in "scikit-learn". A comparison of the experimental results based on *CART* and *Bagging + CART* is shown in Tab. 20, and the accuracy and approximate accuracy iteration curves based on the *Bagging + CART* algorithm and top three attribute groups is shown in Fig. 5.

Table 20 Comparison of experimental results based on Bagging + CART

Classification algorithm	Combination of lexical attributes	Accuracy	Approximate accuracy	*RMSE*	*KC*	*PCC*
CART	All Attributes	33.13%	77.21%	1.3375	0.1954	0.5552
Bagging + CART	All Attributes	36.36%	81.13%	1.2273	0.1789	0.6177
	Top 6 attribute groups based on FS	36.78%	81.90%	1.2318	0.1846	0.6197
	Top 3 attribute groups based on CART	39.67%	85.05	1.1421	0.2120	0.6547



(a) Accuracy Iteration curve (b) Approximate Accuracy Iteration Curve

Figure 5 Experimental Results Based on Bagging + CART Algorithm and Top 3 Attribute Groups

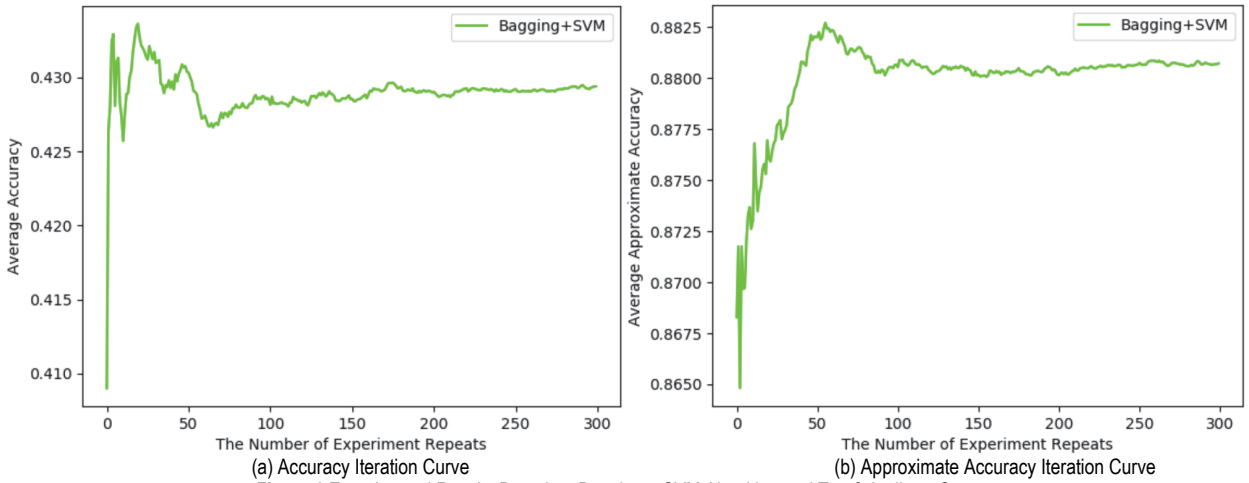
The experimental results in Tab. 20 demonstrate that, compared with the other three classification results in Tab. 20, the accuracy and approximate accuracy based on the Bagging+CART algorithm and top three attribute groups improved accordingly.

(2) Vocabulary grading effect improvement based on the Bagging+SVM integration algorithm

For the SVM classification algorithm, the "gridsearchcv()" function was used to adjust the parameters, and the best combination of parameters was C = 1, gamma = 0.1, kernel = 'rbf'. For the bagging integration algorithm, the "Baggingclassifier()" module in "sklearn.ensemble" in Python was selected. A comparison of the experimental results based on the SVM and Bagging + SVM is shown in Tab. 21, and the accuracy and approximate accuracy iteration curve based on the Bagging + SVM algorithm and top six attribute groups by feature selection is shown in Fig. 6.

Table 21 Comparison of experimental results based on Bagging + SVM

Classification algorithm	Combination of lexical attributes	Accuracy	Approximate accuracy	*RMSE*	*KC*	*PCC*
SVM	All Attributes	42.42%	87.61%	1.0623	0.2307	0.6764
Bagging + SVM	All Attributes	42.51%	87.46%	1.0641	0.2345	0.6801
	Top 6 attribute groups based on FS	42.94%	88.72%	1.0516	0.2352	0.6805



(a) Accuracy Iteration Curve (b) Approximate Accuracy Iteration Curve

Figure 6 Experimental Results Based on Bagging + SVM Algorithm and Top 6 Attribute Groups

Compared with the other two classification results in Tab. 21, the accuracy and approximate accuracy based on the Bagging + SVM algorithm and top six attribute groups improved accordingly. Analyze the reasons, in addition to the advantages of SVM described above, the vocabulary grading model based on "Bagging + SVM," Classification using SVM, and through the sampling method with put back on the original dataset, new datasets are selected to train the classifiers respectively. Thus, the effect of using multiple weak learners to achieve strong learners was achieved, which made the vocabulary classification receive a better effect.

5 CONCLUSION

This paper started with an analysis of the lexical attributes that affect vocabulary grading, followed by an explanation of the extraction of lexical attribute information combined with the constructed word-formation knowledge base, the construction of the mapping functions corresponding to the lexical attributes, and the quantitative representation of the attributes that form the basis for vocabulary grading. Using this as a guide, a vocabulary grading model based on common machine learning classification algorithms was constructed, which included the common machine learning algorithms LR, LDA, KNN, CART, NB, and SVM. In the experiment, the importance of lexical attributes was measured using different methods, and the results demonstrated that the frequency of words in the corpus played an extremely important role. In addition to frequency, the number of semantic items and the average number of strokes of Chinese characters were also important. To improve the effect of vocabulary grading, a variety of feature selection algorithms were used to fuse the importance of lexical attributes on average, then the vocabulary grading experiment was conducted combined with bagging in the integration algorithm. The experimental results demonstrated that the combination of feature selection and the integrated bagging algorithm achieved a better effect. Additionally, because only nine vocabulary attributes were used in the vocabulary grading experiment, this affected the vocabulary grading effect, to a certain extent, and in a follow-up study, we will further explore more lexical attributes to improve the vocabulary grading effect.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 61877004, 62007004); and the Key Project of the National Social Science Foundation of China (No. 18zda295); and the Natural Science Foundation of the Anhui Higher Education Institutions of China (No. KJ2019A0592, KJ2020A0023). We thank Maxine Garcia, PhD, from Liwen Bianji, Edanz Group China (www.liwenbianji.cn/ac) for editing the English text of a draft of this manuscript. We also thank Mo Chen for the constructive advises.

6 REFERENCES

[1] Chomsky, N. (1995). *The Minimalist Program*. MIT Press, Cambridge, Mass.

[2] Wilkins, D. A. (1972). *Linguistics in Language Teaching*. London: Edward Arnold.

[3] Zhang, J., Li, P., Li, Y., Xie, N., & Huang L. (2012). New thinking on the New HSK. *China examinations, 2*, 50-53. https://doi.org/10.19360/j.cnki.11-3303/g4.2012.02.009

[4] Crossley, S. A. & Kristopher, K. (2018). Assessing writing with the tool for the automatic analysis of lexical sophistication (TAALES). *Assessing Writing*. https://doi.org/10.1016/j.asw.2018.06.004

[5] Tracy-Ventura & Nicole (2017). Combining corpora and experimental data to investigate language learning during residence abroad: A study of lexical sophistication. *System*, *71*. https://doi.org/10.1016/j.system.2017.09.022

[6] Higginbotham, G. & Reid, J. (2019). The lexical sophistication of second language learners' academic essays. *Journal of English for Academic Purposes, 37*, 127-140. https://doi.org/10.1016/j.jeap.2018.12.002

[7] Ying, X. (2014). Vocabulary distribution: a measure to words' complexity. *Journal of Yunnan Minzu University, 23*(06), 460-464.

[8] Li, X. (2013). *A comparison of three measures of lexical sophistication*. Nanjing: School of foreign languages and literature, Nanjing University of Technology.

[9] Yuan, P., Chen, Y., Hai, J., & Li, H. (2008). MSVM-kNN: Combining SVM and k-NN for Multi-class Text Classification. *IEEE International Workshop on Semantic Computing and Systems (WSCS)*, 133-140. https://doi.org/10.1109/WSCS.2008.36

[10] Sun, L. B., Yang, J. D., & Yang, H. J. (2006). Hierarchical document categorization with k-nn and concept-based thesauri. *Information Processing & Management, 42*(2), 387-406. https://doi.org/10.1016/j.ipm.2005.04.003

[11] Bo, Y., Xu, Z. B., & Li, C. H. (2008). Latent semantic analysis for text categorization using neural network. *Knowledge-Based Systems, 21*(8), 900-904. https://doi.org/10.1016/j.knosys.2008.03.045

[12] Manikandan, R., Sivakumar, D. R., & In, T. (2018). Machine learning algorithms for text-documents classification: A review. *International Journal of Development Research, 3*(2), 384-389.

[13] Kate, R. J., Luo, X., Patwardhan, S., Franz, M., & Welty, C. (2010). Learning to predict readability using diverse linguistic features. *International Conference on Computational Linguistics*, 546-554.

[14] Schwarm, S. E. & Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. *Association for Computational Linguistics (ACL2005)*, 523-530. https://doi.org/10.3115/1219840.1219905

[15] Sun G. (2015). *Research on readability prediction methods based on linear regression for Chinese documents*. Nanjing: Nanjing University.

[16] Shardlow, M. (2013). A comparison of techniques to automatically identify complex words. *Meeting of the Association for Computational Linguistics student research workshop*, 103-109.

[17] Specia, L., Jauhar, S. K., & Mihalcea, R. (2012). SemEval-2012 task 1: English lexical simplification. *Proceedings of the first joint conference on lexical and computational semantics*, 347-355.

[18] Pantula, M. & Kuppusamy, K. S. (2019). CORDIF: A Machine Learning-Based Approach to Identify Complex Words Using Intra-word Feature Set. *ICoEVCI 2018*. India. https://doi.org/10.1007/978-981-13-1642-5_26

[19] Gala, N., Franois, T., & Fairon, C. (2013). Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *E-lexicography in the 21st century: thinking outside the paper*. Tallin, Estonia. https://doi.org/10.13140/2.1.3913.4089

[20] Gala, N., Franois, T., Bernhard, D., & Fairon, C. (2014). Un modèle pour prédire la complexité lexicale et graduer les mots. *TALN*, 91-102. https://doi.org/10.13140/2.1.4437.6968

[21] Wilkens, R., Vecchia, A. D., Boito, M. Z., Padró, M., & Villavicencio, A. (2014). Size does not matter. Frequency does. A study of features for measuring lexical complexity. *Ibero-American conference on artificial intelligence*, 129-140. https://doi.org/10.1007/978-3-319-12027-0_11

[22] Oregan, J. K. & Jacobs, A. M. (1992). Optimal viewing position effect in word recognition: A challenge to current theory. *Journal of Experimental Psychology: Human Perception and Performance, 18*(1), 185-197. https://doi.org/10.1037/0096-1523.18.1.185

[23] Xing, H. (2005). A statistics analysis of components of the character entries in the HSK Graded Character List. *Chinese Teaching in the World, 72*, 49-55.
[24] Guo, D. (2016). *Analyzing on dynamic words and their structural modes in building the sentence-based treebank*. Beijing: The college of information and technology, Beijing Normal University.
[25] Guo D., Zhu, S., et al. (2016). Construction of the dynamic word structural mode knowledge base for the international Chinese teaching. *Proceedings of the 16th Chinese Lexical Semantics Workshop (CLSW2016)*. https://doi.org/10.1007/978-3-319-49508-8_24
[26] Zhang, Y., Song, J., Peng, W., Guo D., & Zhang J. (2019). Construction and analysis of Chinese word-formation knowledge base based on Modern Chinese Dictionary. *Proceedings of the 20th Chinese Lexical Semantics Workshop (CLSW2019)*. https://doi.org/10.1007/978-981-32-9240-6_16
[27] Crossley, S. A., Cobb, T., & Mcnamara, D. S. (2013). Comparing count-based and band-based indices of word frequency: Im- plications for active vocabulary research and pedagogical applications. *System, 41*, 965-981. https://doi.org/10.1016/j.system.2013.08.002
[28] Li, H. (2019). *Statistical learning methods* (Second Edition). Beijing: Tsinghua University Press.
[29] Yang, Y. & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. *ICML, 97*, 412-420.
[30] Zhang, Y., Song, J., Peng, W., Guo D., & Song T. (2020). Quantitative Analysis of Chinese vocabulary comprehensive complexity based on AHP. *Journal of Chinese Information Processing, 34*(12), 17-29.

Contact information:

Yinbing ZHANG, PhD student
1) School of Artificial Intelligence,
Beijing Normal University,
No. 19, XinJieKouWai St., HaiDian District, Beijing 100875, P. R. China
2) School of Mathematical Science,
Huaibei Normal University,
No. 100, Dongshan Road, Huaibei, Anhui 235000, P. R. China
E-mail: zhangyinbing@mail.bnu.edu.cn

Jihua SONG, PhD, Professor
(Corresponding author)
School of Artificial Intelligence,
Beijing Normal University,
No. 19, XinJieKouWai St., HaiDian District, Beijing 100875, P. R. China
E-mail: songjh@bnu.edu.cn

Weiming PENG, PhD
(Corresponding author)
School of Artificial Intelligence,
Beijing Normal University,
No. 19, XinJieKouWai St., HaiDian District, Beijing 100875, P. R. China
E-mail: pengweiming@bnu.edu.cn

Dongdong GUO, PhD student
School of Artificial Intelligence,
Beijing Normal University,
No. 19, XinJieKouWai St., HaiDian District, Beijing 100875, P. R. China
E-mail: dongdongguo@mail.bnu.edu.cn

Tianbao SONG, PhD
School of Computer Science and Engineering,
Beijing Technology and Business University,
No. 11, Fucheng Road., HaiDian District, Beijing 100048, P. R. China
E-mail: songtianbao@btbu.edu.cn