ISSN 2522-6398 Vol. 5, Issue 6: 93-96, DOI: 10.25236/FER.2022.050618

Research on English Text Difficult Auditing Based on Artificial Neural Network

Hang Xu[1,†], Yi Yu[2,†], Shuai Li[1,*]

[1]*Nanchang Institute of Technology, Nanchang, Jiangxi, China*
[2]*Macau University of Science and Technology, Macau, China*
[*]*Corresponding author*
[†]*These authors contributed equally*

Abstract: *English text difficulty classification is very important in all walks of life, so it is very important to determine the text complexity of an article quickly and accurately. We select volumes 1 to 4 of New English Concept as a sample, in order to make up for the lack of terminology and applicability in the New Concept textbook, reader's Digest USA, Time and Nature, CET-4, CET-6 and IELTS are selected as the supplementary samples. The LCA method and L2SCA method were used to extract 33 lexical-related and 23 syntactic-related indicators from the sample articles as the original data set. The data were input into the artificial neural network for training and cross-validation, the training effect was excellent. After the above processing, the remaining five kinds of sample data were substituted into the trained artificial neural network for text complexity assessment. The standardized reading difficulty coefficients obtained from the assessment were drawn and the results of all kinds of data were found to be in line with the recognized difficulty ranking.*

Keywords: *Artificial Neural Network; Hierarchical Clustering*

1. Introduction

From the aspect of reading, it can be seen that text complexity directly affects the readability of an article and is an important indicator for understanding an article. Textual complexity has two levels: lexical and syntactic. Lennon, in 1990, defined the complexity "A large number of words and structures are used in articles", wolfe-Quintero also published in 1998. [1] Complexity includes both lexical and syntactic complexity. Now, in 2021, it's generally accepted. The complexity of words refers to the diversity of bases and complexity of words. Syntactic complexity refers to the existence of a large number of things in a language It contains both syntactic complexity and morphological complexity. Complex syntax Sex is an important component of syntactic complexity. [2]

2. Text Difficult Judge Model Based on Artificial Neural Network

Vocabulary is the most basic element of language, the analysis of vocabulary is not to calculate the number of words in the article, but to analyze the diversity and complexity of vocabulary. Lexical complexity reflects the depth and breadth of an article. So the exploration of the complexity of the article must be the exploration of the lexical complexity.

The lexical complexity is analyzed using the LCA. Lexical Complexity Analyzer uses lexical complexity index to analyze vocabulary. The LCA table includes 33 indexes, such as word density (LD), word advanced degree (LS1), verb advanced degree (VS1), correction type ratio (CTTR), different word number (NDW) and correction TTR (CTTR). [3]

In an article, words are the basic individuals of the article, while sentences are the veins of the article.

In analyzing syntactic complexity, the commonly used analysis methods mainly include manual annotation and automatic annotation. People, in the process of labeling, it is subjective and time-consuming. With the rapid development of The Times today. In analyzing syntactic complexity, the commonly used analysis methods mainly include manual annotation and automatic annotation. People. In the process of labeling, it is subjective and time-consuming. With the rapid development of science and technology. Compared with manual labeling, computer program labeling has the advantages of speed, objectivity, accuracy and reliability. When choosing a syntactic analysis scheme, the search for accurate

ISSN 2522-6398 Vol. 5, Issue 6: 93-96, DOI: 10.25236/FER.2022.050618

machine search becomes a key factor. When choosing an analyzer that addresses syntactic complexity in text complexity, we use L2SCA to analyze the number of times the selected metric occurs in the article.

We measured the Syntactic Complexity using L2SCA (L2 Syntactic Complexity Analyzer) Index information table, obtained indicators of 5 categories: word length, word density, sentence complexity parallel Degree, phrase structure. The common indicators are W/S (average sentence length), W/T (average unit length of T),W/C (average clause length), C/T (clause to T unit ratio), CT/T (complex T unit to T unit ratio),DC/C (clause to clause ratio), DC/T (clause to T unit ratio), C/S (clause to sentence frequency ratio),CP/C, CP/T, T/S (T units and sentences), CN/C (ratio of complex NOMINAL phrases to clauses), CN/T (ratio of complex nominal phrases to clauses),VP/T (ratio of verb phrase to T unit), 23 in total. [4]

In order to conduct quantitative analysis for articles of different levels, the complexity of tested articles should have accuracy, applicability and universality. Training in more representative articles can improve the performance of the model. So selected text teaching materials must be in the life the coverage of the larger series of articles, reference consulting experts and literature in many aspects, on the basis of our proposed the new concept of teaching materials in four copies, "college English teaching material", "Reading", "Time", "Nature", "IELTS", six sets of domestic comparing graded teaching material inspection, to determine the points

On the basis of reasonable level, each set of teaching materials selected 40 articles as neural network learning and testing. Among them, 80% of the selected articles were randomly selected as the training set for nerve. Network training; The remaining 20% serves as a test set for verifying the complexity of the text.

Artificial neural network (ANN) is developed from perceptron. Its network learning mode adopts supervised learning, which is usually composed of one input layer, one or more hidden layers and one output layer. The construction of a complete process consists of a forward propagation process and a back propagation process. Using the index of principal component dimension reduction analysis, the training set articles are input into the neural network model.

The pre-processed data set was randomly sampled, 80% of which was used as the training set and the remaining 20% as the test set. Factor analysis is used to extract common factors from indicators, and the extracted common factors are used as input variables of neural network. The hidden layer and the number of neurons need to be analyzed and discussed.[5]

In forward propagation, the output value of node of hidden layer is set as:

$$a_j = f\left(\sum_{i=1}^{m} w_{ij}\, x_{ij} + \theta_{ij}\right)$$

$$f(x) = \frac{1}{1 + e^{-x}}$$

Where, the node output layer of hidden layer is a_j; The activation function is $f(\cdot)$; The eigenvector of the input layer is x_{ij}; The weights and thresholds from input layer to output layer are w_{ij} and θ_{ij} respectively.

After the output value of the hidden layer is obtained, it is calculated as the input of the output layer. Finally, the output layer result expression is as follows:

$$y_k = f(a_k) = f\left(\sum_{j-1}^{n} w_{ij}\, x_{ij+}\theta_{jk}\right)$$

In the back propagation learning, the expected and output signals of neural network training are t_k by means of simulation layer data and model verification. The actual output signal of the neural network is y_k. Then the error function E is:

$$E = \frac{1}{2}\sum_{k=1}^{n}(y_k - t_k)^2$$

As mentioned above, for the accuracy of artificial neural network simulation, error function E is used to ensure that the error between y_k and t_k is reduced to a given accuracy range of 0.1. Gradient descent

ISSN 2522-6398 Vol. 5, Issue 6: 93-96, DOI: 10.25236/FER.2022.050618

method is used for error function E to update weight w_{jk} and w_{ij} of hidden layer and output value respectively. The formula is as follows:

$$w_{jk} = w_{jk} - \eta\delta_o(k)a_i$$

$$w_{ij} = w_{ij} - \eta a_j(1-a_j)\left(\sum_{k=1}^{N}\delta_o(k)w_{jk}\right)a_i$$

When the error is reduced to a given precision, the training can end. The above can obtain the specific results of the artificial neural network model.

At the end of the training, the mean variance and the mean absolute value are output. The smaller the mean variance and the mean absolute value, the better the learning degree of neural network.

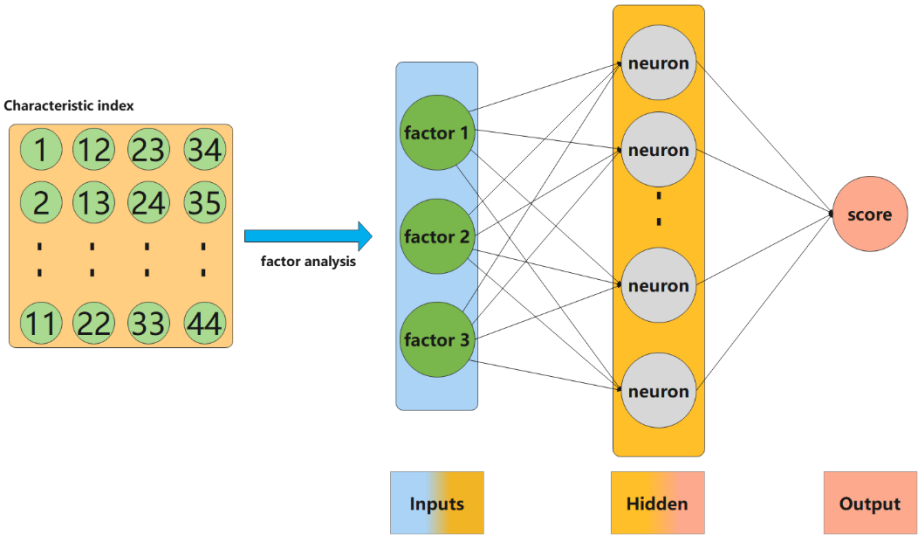

Figure 1: Artificial neural network for text complexity model

First, we analyze the three factors: the number and type of complex words, sentence phrasal verbs, and the type and number of words. Based on hierarchical clustering, five kinds of articles of New Concept English 1, New Concept English 2, New Concept English 3, New Concept English 4 and Nature are hierarchical clustering. Analysis factor to article classification is consistent with the official website information. The hierarchical clustering analysis diagram is as follows:

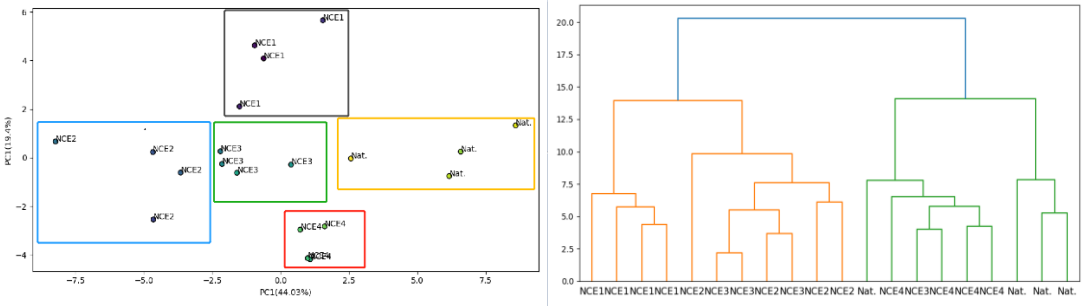

Figure 2: Hierarchical clustering diagram of difficulty

Through the above hierarchical cluster analysis, the feasibility of the five factors, such as the number and type of complex words, phrasal verbs of sentences, and the type and number of words, is illustrated.

They are graded for CET4, CET, IETLS, TIME and READ. 40 articles in each category were randomly selected to make a broken line statistical graph and observe the reading difficulty coefficient scored by the analysis model.

ISSN 2522-6398 Vol. 5, Issue 6: 93-96, DOI: 10.25236/FER.2022.050618

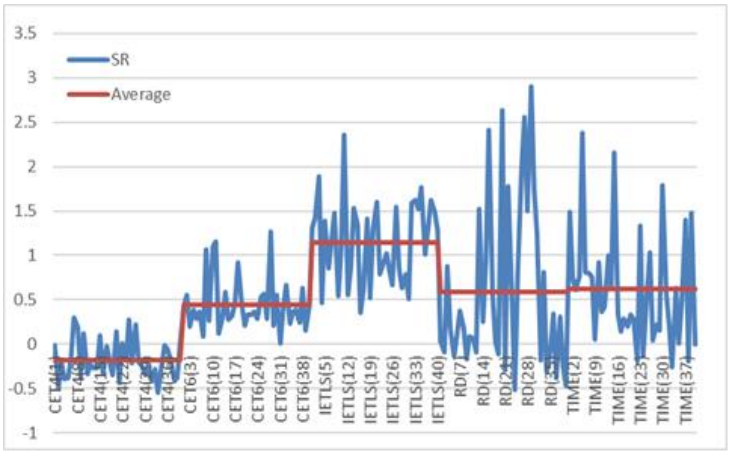

Figure 3: Neural network score

As can be seen from above, the blue line is the score line graph of the neural network for the article, and the red line is the average score of all kinds of articles. From the ordinate view, the standardized reading difficulty coefficients defined by us are basically between (-1~3). From the abscissa view, CET-4 articles are basically below the coordinate 0, CET-6 articles float around 0.5, and IETLS articles average 1.5. For magazine articles: TIME articles (TIME articles), reader articles (RD), both similar to CET-4 articles, low level, but also has a higher degree of difficulty. This is also in line with the universality of scientific magazine articles such as TIME articles and Reader articles, which are suitable for all social strata. And like English cet-4 article chapter, English cet-6 article, IELTS article and other subject nature, are basically the level of English examination article, so, they are clustered at the same level by area comparison. All kinds of data predicted results basically accord with the recognized difficulty ranking.

3. Conclusion

Based on LCA and L2SCA, this paper summarizes and analyzes the traditional research on text difficulty, and uses ANN to study the measurement model of English text legibility from lexical and syntactic aspects. The research of the measurement model of the measurement accuracy, applicable scope, etc., are more than the measurement model of traditional article legibility and has strong theoretical significance and use value.

References

[1] Yan Shijuan. Eye movement study on the influence of text difficulty and word segmentation training on college students' text reading effect [D]. ludong university, 2021.
[2] Wu Mengyao. Research on the difficulty of textual English teaching materials based on corpus [D]. Northern National University, 2021.
[3] Zhao Yuxuan, Xu Jianfen. Chinese English Major Student Academic Papers (English) [J]. Foreign Language Education, 2020, 20 (00): 17-29.
[4] Wei Yan, Cai Yany. SP written language syntax complexity - the quantitative study of L2SCA syntactic analysis tool [J]. Journal of Wuyi University, 2018, 37 (10): 52-56.
[5] Zhu Qixun. The text sequence analysis method using deep learning and its emotional deduction [D]. University of Electronic Science and Technology, 2020.