International Journal of Social Science and Education Research Volume 5 Issue 4, 2022
ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202204_5(4).0020

Study on the Difficulty Evaluation of English Reading Text Based on Entropy-Topsis Model

Hongxia Ye*, Feifei Li, Xing Wang, Junjie Tao

Anhui University of Finance and Economics, Bengbu City, Anhui Province, 233030, China

Abstract

A study on judging the difficulty of reading texts in English. We first search for data, collect data and preprocess the data. We choose TOPSIS algorithm with entropy weight to judge the difficulty of cet-4 reading text. We make corresponding assumptions and select indicators, and use entropy weight -TOPSIS algorithm to calculate the score. We use MATLAB to calculate the results, and rank the reading difficulty of cet-4 papers in different years. At the same time, we compared the results with the average score of CET-4 in the same year. If the test papers with higher default scores are easier to read, and those with lower scores are more difficult to read, the change trend of English reading difficulty reflected by the two is basically consistent.In order to test the accuracy of the model, we also select the relevant data of CET-6 test paper, and use our model to calculate the score. Compared with the average score of CET-6 test paper of the same year, we find that its changing trend is consistent with the score we calculated.

Keywords

Reading Difficulty Judgment; Entropy weight TOPSIS algorithm.

1. Research Background

At present, English is one of the required courses for students in our country, among which reading is one of the four English cores. English reading texts occupy a large proportion in all kinds of textbooks and play a vital role in English learning. Therefore, the English reading difficulty test has been attached great importance by the Research institute of English Education in China in recent years. The ability to read and understand English texts is an important ability for us to learn English. Meanwhile, English reading is also the main way for us to understand the world and obtain information.

In the process of learning English, the richness of knowledge we acquire will have some influence on our reading comprehension level. Meanwhile, our accumulated Knowledge of English cultural background plays a decisive role in our speed of English reading. But at present, Our country lacks a good English language communication environment, let alone some necessary cultural values. Therefore, for English learners, the acquisition of English language knowledge is inseparable from daily English reading, and the difficulty of reading English texts is directly related to our interest in reading and motivation to insist on reading. The accumulation of English knowledge in many aspects, reading English periodicals, magazines and so on will enable us to have better English reading ability.

The control of the difficulty of English texts can help English learners at different stages to find articles that are more suitable for their reading levels, so that they can get a sense of achievement and challenge themselves to read English texts with higher difficulty. The analysis of the influencing factors of English reading difficulty has a certain guiding significance for us to further improve our ability of English reading .

Entropy weight method is an objective decision - making method. According to the basic principles of information theory, information is a measure of the degree of order of a system

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202204_5(4).0020

[1]. According to the definition of information entropy, entropy value can be used to judge the dispersion degree of an index. The smaller the information entropy is, the greater the dispersion degree of the index is, and the greater the influence (i.e. weight) on the comprehensive evaluation is. If all the values of this index are equal, it does not play a role in the comprehensive evaluation. Therefore, the weight of each index can be calculated by using the tool of information entropy to provide a basis for comprehensive evaluation of multiple indexes [2].

TOPSIS algorithm is a sequential optimization technique of ideal objective similarity, and it is a very effective method in multi-objective decision analysis. Through the normalization matrix of normalized data, it finds out the optimal target and the worst target [3] (represented by ideal solution and anti-ideal solution respectively), calculates the distance between each evaluation target and the ideal solution and anti-ideal solution respectively [4], obtains the closeness degree of each target and the ideal solution and sorts them, which serves as the basis for evaluating the quality of the target. The value of proximity is between 0 and 1. The closer to 1, the closer to the optimal level, and the closer to 0, the closer to the worst level [5]. This method has been successfully applied in many fields such as land use planning, material selection and evaluation, project investment and so on, which obviously improves the scientific, accuracy and maneuverability of multi-objective decision analysis.

2. Research Idea

This paper mainly establishes a model for the evaluation of English reading difficulty. If we only consider the average length of sentences and the average number of syllables of words, it is too one-sided, and the error is large, so we can't accurately judge the difficulty of an English text. We consider using TOPSIS Model Based on entropy weight method. For the convenience of calculation, we take CET-4 and CET-6 English reading text as an example. We use the number of sentences and words of CET-4 and CET-6 English reading texts in recent years as data, select the average length of sentences, the familiarity of words and the average number of characters of words as indicators, calculate the weight by entropy weight method, and then use TOPSIS algorithm to score according to the relative closeness to judge the difficulty of English reading texts.

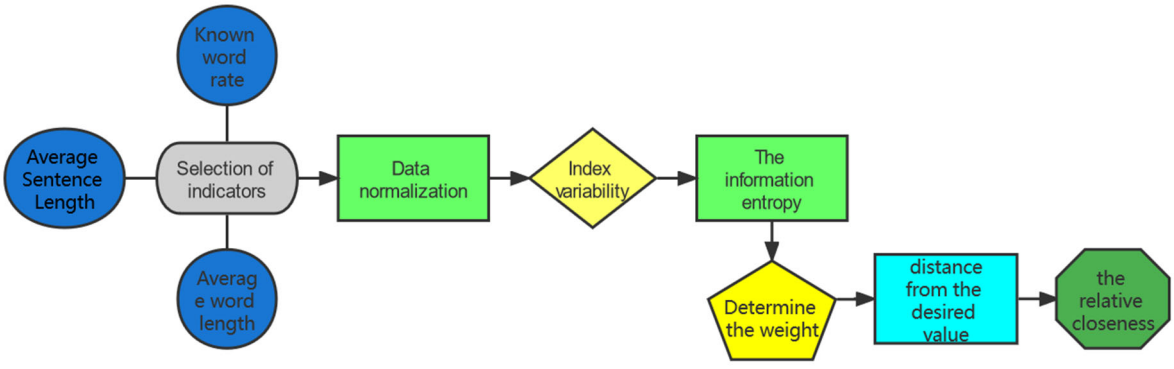

Figure 1. The Modeling Process Thinking Diagram

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202204_5(4).0020

3. The Research Methods

3.1. Selection of Indicators

3.1.1. The Selection of Average Sentence Length Index

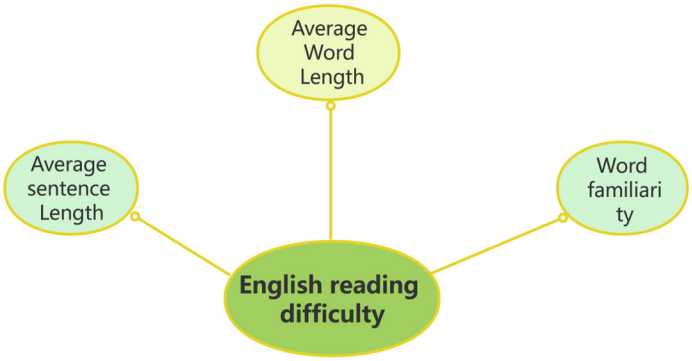

Figure 2. The Selection of Three Indicators

We use the ratio of the total number of words to the total number of sentences in the CET-4 paper each year as the average sentence length. The average length of a sentence reflects its complexity. The longer the average length of a sentence is, the more complex the sentence is and the more difficult it is for readers to understand. Correspondingly, the shorter the average sentence length, the less complex it is, and relatively easier to understand.Its calculation formula is as follows:

$$\lambda_i = \frac{b_i}{s_i}(i = 1,2,3,4,5)$$

3.1.2. The Selection of Known Vocabulary Rate Index

The ratio of known words is the ratio of the number of known words to the total number of words. The higher the proportion of known words in the total words is, and the text is easier to read. Conversely, The lower the proportion of known words in the total words is, the more difficult it is to read.Its calculation formula is as follows:

$$\lambda_2 = \frac{b_i - m_i}{b_i}(i = 1,2,3,4,5)$$

3.1.3. The Selection of Index of Average Character Length of Words

Complex words can cause a certain amount of reading pressure on readers. We take the ratio of the total number of characters to the total number of words in CET-4 as the average character length of words. The longer the average character of a word is, the more complex the word is and the more difficult it is to read. If most of the words are simple, it will be easier for readers. Its calculation formula is as follows:

$$\lambda_3 = \frac{a_i}{\mathrm{b}_i}(i = 1,2,3,4,5)$$

3.2. Data Normalization

Because the measurement units of each index are not unified, they should be normalized before calculating the weight, that is, the absolute value of the index is converted into relative value. At the same time, there are both positive and negative indicators in the data used in this paper. In order to prevent 0 in both positive and negative indicators, we need to improve the normalization formula of the general evaluation model.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202204_5(4).0020

For positive indicators:

$$\lambda_{ij} = 0.98\frac{\lambda_{ij} - \min(\lambda_{ij})}{\max(\lambda_{ij}) - \min(\lambda_{ij})} + 0.02$$

For negative indicators:

$$\lambda_{ij} = 0.98\frac{\max \lambda_{ij}}{\max \lambda_{ij} - \min \lambda_{ij}} + 0.02$$

3.3. Calculate the Variability of Indicators

We need to determine the weight according to the variability of the index. Its calculation formula is as follows:

$$e_j = -\frac{\sum_{i=1}^{5} p_{ij} \ln p_{ij}}{\ln 5}$$

3.4. Calculate the Information Entropy of Each Indicator

Information entropy refers to the expectation of the amount of information, and we can understand it as uncertainty. Generally speaking, the smaller the information entropy of an index is, the greater the variation degree of the index is, the more information it provides and the more important it plays in the comprehensive evaluation. Its calculation formula is as follows:

$$e_j = -\frac{\sum_{i=1}^{5} p_{ij} \ln p_{ij}}{\ln 5}$$

3.5. Determine the Weight of Each Indicator

If the information entropy of each indicator is ej, we can obtain the weight of each indicator by information entropy calculation[6]:

$$w_{ij} = \frac{1 - e_i}{n - \sum e_i}$$

3.6. Calculate the Distance from the Desired Value

The positive ideal solution is:

$$\lambda^{+}{}_j = \max_{1 \le i \le 5} \{\lambda_{ij}\}$$

The negative ideal solution is:

$$\lambda_j^{-} = \min_{1 \le i \le 5} \{\lambda_{ij}\}$$

Then the distance is:

$$D_i^{+} = \sqrt{\sum_{j=1}^{3} \left[w_i (\lambda_j^{+} - \lambda_{ij})^2\right]}$$

$$D_i^{-} = \sqrt{\sum_{j=1}^{3} \left[w_i (\lambda_j^{-} - \lambda_{ij})^2\right]}$$

4. Results Analysis

Using MATLAB, the weights of the three indicators are 0.3944, 0.3272, 0.2784 in turn

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202204_5(4).0020

From 2007 to 2011, the scores of difficulty in reading papers were 0.3407, 0.2038, 0.1915, 0.1629, 0.1011

The result is shown below:

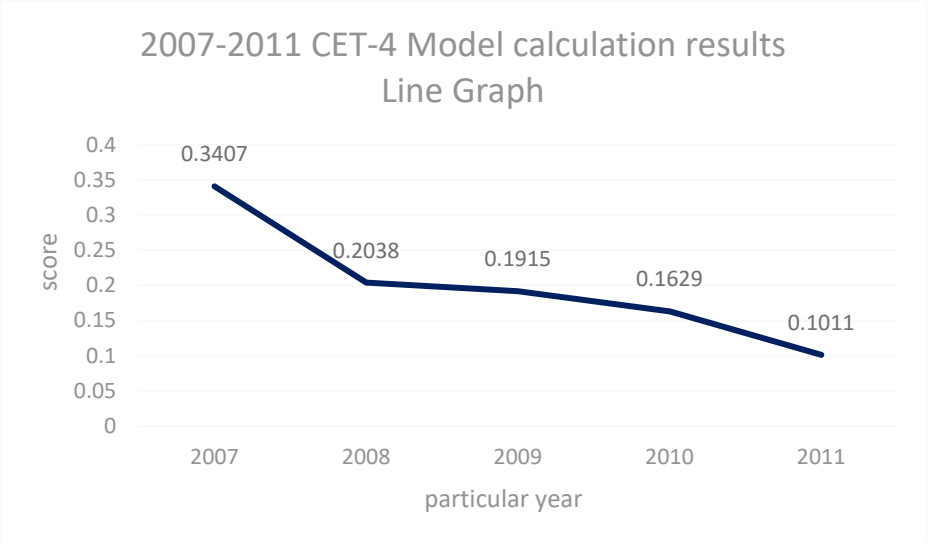

Figure 3. 2007-2011Model Calculation Results

We can get results from the model that the highest score is in 2007, and it's the closest to the ideal value, that is, it is easier for readers to read. While the lowest score is in 2011, it is more difficult to read the paper in that year. Our comparison with the CET-4 average grade of the same year shows that the CET-4 average grade was the highest in 2007 and the CET-4 average was the lowest in 2011, with similar change trend.Therefore, the model established by this method can accurately judge the degree of difficulty in reading English texts.

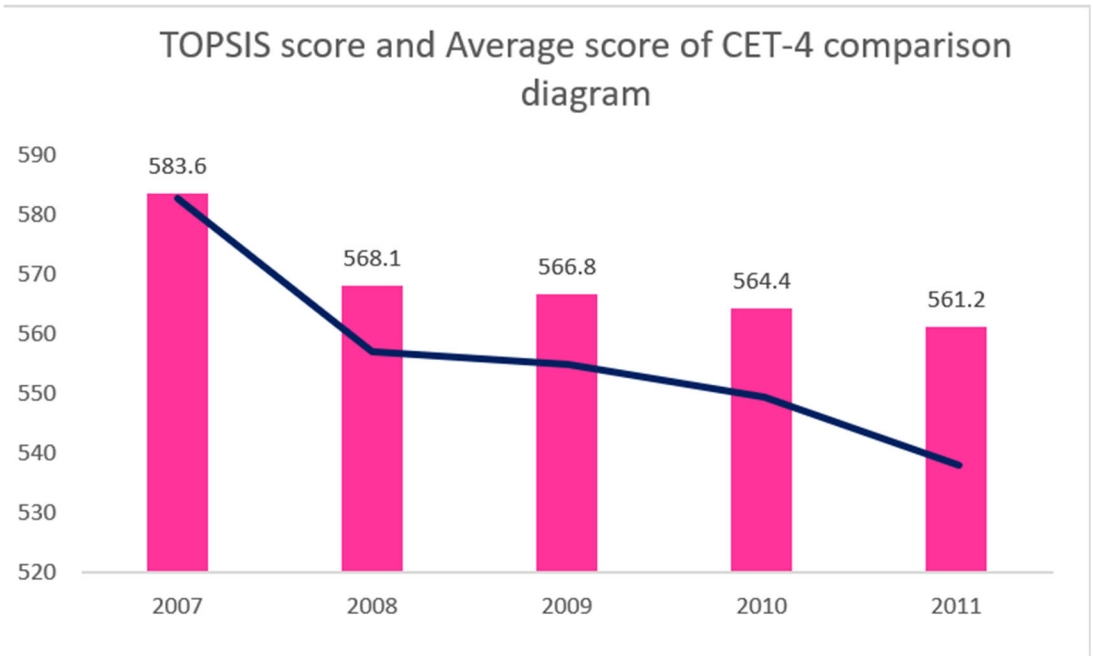

Figure 4. A Comparison between the Difficulty of Our Calculation and the Score of CET-4 Grade in the same year

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202204_5(4).0020

Annotation:
The broken line graph shows the changing trend of the final score using TOPSIS algorithm
The histogram shows the average score of CET-4 from 2007 to 2011

5. Results of Inspection

In order to verify the correctness of the model, we collected the relevant data of cet-6 test papers in recent years, and calculated the difficulty score of this year's test paper by using the reading difficulty test model, and compared it with the average score of CET-6 test papers in the same year. We processed the data as follows

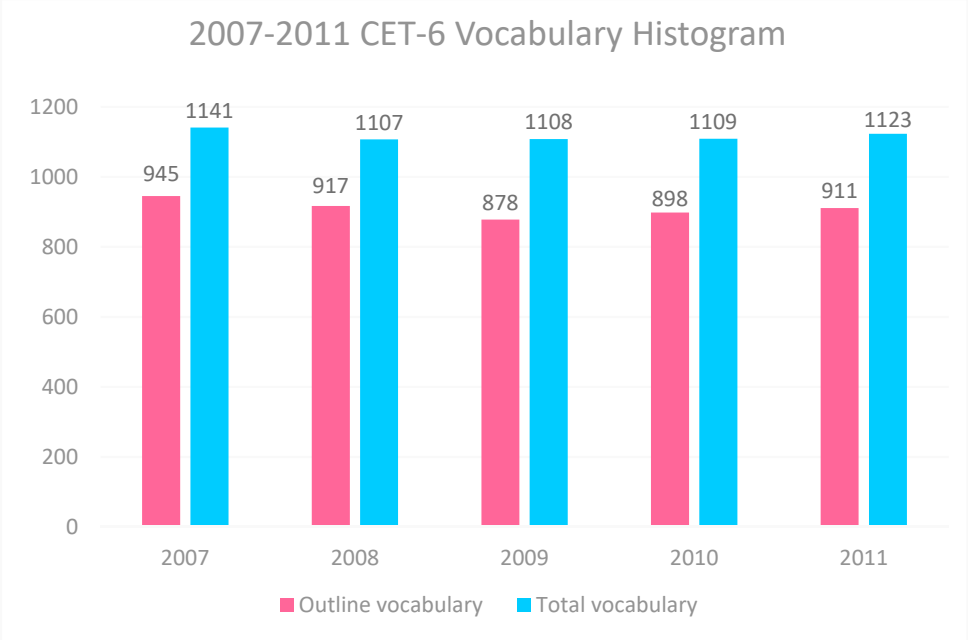

Figure 5. 2007-2011 Histogram of CET-6 Vocabulary

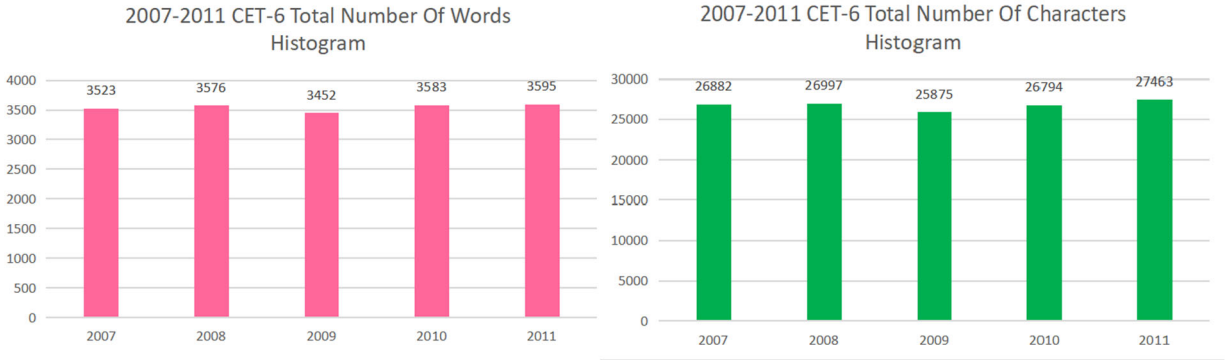

Figure 6. 2007-2011 Total Number of Words and Characters in CET-6

The result is shown below. We found that the results we calculated were similar to the average score of cet-6 in the same year, which verified the accuracy of our model.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202204_5(4).0020

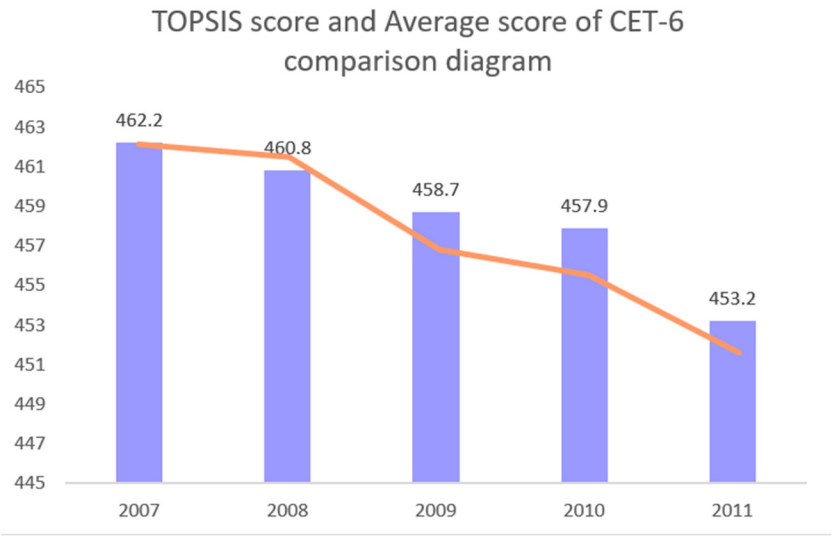

Figure 7. A Comparison between the Difficulty of our Calculation and the Score of CET-4 Grade in the same year

6. Conclusion

We select the average sentence length, the ratio of known words and the average character length of words as indicators, and use the entropy-Topsis model to score the English reading text and judge its difficulty. For readers, this can facilitate readers to choose the text, in order to achieve better reading effect. For teachers, this is convenient for teachers to teach students according to their aptitude, so as to achieve better teaching results. At the same time, for different languages, we only need to change the selection of indicators and use this model to judge the degree of difficulty, which provides convenience for us to learn foreign languages.

References

[1] Chen Chong. Evaluating the Investment Benefit of Multinational Enterprises' International Projects Based on Risk Adjustment: Evidence from China[J]. EURASIA Journal of Mathematics, Science and Technology Education,2016,12(9).

[2] Zhi Qiang Zhao,Zhi Gang Wang. The Application of Entropy Weighting Ideal Point Method in Electronic Information Equipment System of Systems Construction Decision-Making[J]. Applied Mechanics and Materials,2013,2601(385-386).

[3] Chen Jiang,Qiao Li Mi. Research on the Grey Correlation Evaluation Model of Equipment Utilization Quality Based on TOPSIS[J]. Applied Mechanics and Materials,2013,2733(438-439):

[4] Aleksandra Schwenk-Ferrero,Andrei Andrianov. Comparison and Screening of Nuclear Fuel Cycle Options in View of Sustainable Performance and Waste Management[J]. Sustainability,2017,9(9).

[5] Yuan Ronglian,Ai Mingye,Jia Qiaona,Liu Yuxuan. Evaluation index system of steel industry sustainable development based on entropy method and topsis method[J]. IOP Conference Series: Earth and Environmental Science,2018,128(1).

[6] Jin Hui Liu,Ling Kang Chen,Chuan Yi Liu,Lan Rong Qiu,Shu He. Pb speciation in rare earth minerals and use of entropy and fuzzy clustering methods to assess the migration capacity of Pb during mining activities[J]. Ecotoxicology and Environmental Safety,2018,165.

[7] Zhu Li Wu. Analysis on Differential Equation of Decision Model Based on Matlab Simulation[J]. Applied Mechanics and Materials,2014,3365(602-605).