Research on AI-based English Vocabulary Statistics

Lili Lu
Jiangsu College of Finance & Accounting, Lianyungang 222067, China

Abstract
Artificial intelligence ("AI") studies the law of intelligent activities of human beings. In college English teaching, AI technology and computer-aided teaching system can provide statistics of English vocabulary and realize sharing of teaching resources for higher teaching quality. In view of existing problems with college English teaching, such as single teaching model and shortage of teachers, the paper conducts research on and design of an intelligent computer assisted system based on AI technology. Then, through application of AI technology and based on LEXSTAT program, the paper counts and analyzes English words in the computer system, and studies the distribution rules of some words, with a view to realize intelligent statistical analysis of English teaching, help teachers do a better job in teaching and enhance students' ability to learn independently.

Keywords: Artificial intelligence, Computer, Aided, English vocabulary, Statistics.

1. RESEARCH BACKGROUND
1.1 Literature review
With development of computer technology, artificial intelligence ("AI") has been widely used in educational technology and gradually become a major trend in computer-aided teaching (Wan and Guan, 2006). In statistical survey, combination of mutual information and information entropy is more effective than mutual information alone. With regard to filtering method of language chunk boundary rules, chunk left and right boundary rules and stop words exert a greater influence on the chunk extraction result. According to research results, in machine-aided translation system, extracting the repetitive language units with existing chunk extraction method makes it easier to construct the translation memory bank and in turn improves the translation efficiency (Jiang et al., 2007). In addition, with improvement in educational technology, AI has seen wider application in teaching. For example, improving the teaching model as per intelligent computer assisted instruction ("ICAI") principle can significantly enhance the teaching effect (Liu et al., 2010). Intelligent computer can simulate the memory and thinking of human beings, so computer-aided teaching system can help achieve intelligent teaching, solve the problems with imagery thinking and insightful thinking, and rapidly promote the independent learning and expressive ability of students (Zhao et al, 2011).

1.2 Research purpose
In modern education, computer-aided teaching has been widely used, but there are still some problems with it. For example, there are strict restrictions on the application time and space, and the courseware is of low flexibility (Xu and Hu, 2011). With advancement of technologies in the information age, AI has emerged and been gradually popularized, with the number of applications on the increase. Computer & Internet technology and ICAI system can effectively make up for the shortcomings of traditional computer-aided teaching and make real intelligent teaching a reality. The application of AI technology in education and teaching can facilitate growth of students and teachers and promote reforms in education (Li and Shen, 2007). Moreover, with introduction of AI technology, computer-aided teaching is being transformed into ICAI. ICAI system has made up for deficiencies of traditional computer-aided teaching system, and significantly improved teaching effect and efficiency (Wang et al., 2003). From the perspective of module function and key technologies, the ICAI system shall be based on the Web, including knowledge base, student model and teacher model (Zhang and Li, 2006). Moreover, it can be used not only in colleges, but also in high schools, middle schools and so on. By following the principles of improving system and deepening teaching integration, AI technology is adopted to improve the English teaching system and to create a comprehensive model and environment for independent language learning (Chen et al., 2008).
2. OVERVIEW OF AI

AI is a borderline science emerging due to interpenetration of many disciplines, and mainly involves computer science, information theory, cybernetics, linguistics, neurophysiology and other disciplines. Based on the theory of natural intelligence, AI simulates the awareness and thinking of human beings through simulation, extension and expansion. At present, AI research focuses on 3 hot topics, namely, intelligent interface, data mining, and the subject system. Intelligent interface technology can promote communication between people and computers. With knowledge representation method, AI technology uses computers to realize many practical functions, such as text expression, translation in different languages, speech recognition, image recognition, and natural language understanding, thus having great application value and basic theoretical significance (Zeng et al., 2006). The actual application data is fuzzy, incomplete and numerous, but potentially useful information can be extracted through technical means, which is called data mining. By far, three strong technical pillars have been formed: database, AI and mathematical statistics. The main research contents include: knowledge representation method, discovery algorithm, visualization technology, data warehouse, qualitative and quantitative exchange model, etc. The subject system refers to the entity with mental state, including the intention, ability, belief, choice and so on. Featured by intelligence and autonomy, the subject system can accomplish tasks in an autonomous and independent manner, interact with the environment and communicate with other subjects to achieve the goal (Ma, 2012). The current research on subjects and multi-subject systems mainly focuses on the subject’s architecture and organization, coordination among subjects, multi-subject learning and multi-subject system application, etc.

3. SYSTEM ANALYSIS AND DESIGN

The paper only involves two majors, i.e. computer science and machine tooling, and researches the total number of words, vocabulary list and coverage rate. The statistics is done based on word form and different forms of words are counted as a separate item.

3.1 Total number of words

Based on searching and statistics of English teaching materials for the two majors, nearly 3, 500 words are collected from the computer science major and about 3, 000 words collected from the machine tooling major, as is shown in Table 1.

<table>
<thead>
<tr>
<th>Major</th>
<th>No. of characters</th>
<th>No. of words</th>
<th>Vocabulary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computer science</td>
<td>18,131</td>
<td>3,486</td>
<td>317</td>
</tr>
<tr>
<td>Machine tooling</td>
<td>15,452</td>
<td>3,022</td>
<td>275</td>
</tr>
</tbody>
</table>

For the convenience of further research, only English proper nouns are included in the total number of words, excluding mathematical symbols and other scientific identifiers (Li et al., 2008). Moreover, punctuations aren’t included in the total number of characters. Then, if the total number of characters is represented by $L$ and the total number of words by $W$, then the average word length $\bar{L}_W$ is:

$$\bar{L}_W = \frac{L}{W}$$  \hspace{2cm} (1)

$$\bar{L}_W(\text{computer}) = \frac{18131}{3486} \approx 5.201$$  \hspace{2cm} (2)

$$\bar{L}_W(\text{machine tool}) = \frac{15452}{3022} \approx 5.113$$  \hspace{2cm} (3)

3.2 Vocabulary list

Computer is employed to work out the positive sequence vocabulary list and frequency vocabulary list of the two majors. As the list is rather long, only a part of the list is displayed. Vocabulary frequency and text distribution rate are the major elements considered when preparing the list of frequently used English words for science and technology. Another key element considered is major-specific distribution of words, that is,
distribution of the same word in different majors, as is shown in Table 2. Take the computer science major as an example. Table 2 shows the number of occurrence and the number of sample units of the selected word. Then, it is possible to calculate the frequency of each word and text distribution rate.

Table 2 Vocabulary List of Computer Science Major

<table>
<thead>
<tr>
<th>Word</th>
<th>No. of occurrence</th>
<th>No. of sample units</th>
<th>Word</th>
<th>No. of occurrence</th>
<th>No. of sample units</th>
</tr>
</thead>
<tbody>
<tr>
<td>of</td>
<td>1,376</td>
<td>147</td>
<td>is</td>
<td>1,751</td>
<td>148</td>
</tr>
<tr>
<td>to</td>
<td>1,876</td>
<td>153</td>
<td>be</td>
<td>986</td>
<td>145</td>
</tr>
<tr>
<td>and</td>
<td>1,844</td>
<td>153</td>
<td>the</td>
<td>3,874</td>
<td>155</td>
</tr>
<tr>
<td>a</td>
<td>1,065</td>
<td>154</td>
<td>in</td>
<td>2,011</td>
<td>147</td>
</tr>
</tbody>
</table>

Take the word “accord” as an example, whose frequency is:

\[
F(\text{accord}) = \frac{8}{3486} = 0.00229
\]  

(4)

And whose text distribution rate is:

\[
Dt(\text{accord}) = \frac{38}{155} = 0.24516
\]  

(5)

3.3 Coverage rate

Coverage rate \( C \) refers to the proportion of the total occurrence number of some words to the total number of words \( W \) (He and Wang, 2008). The formula for calculating the coverage rate is:

\[
C = \sum_{i=1}^{n} F_i
\]  

(6)

where, the frequency of the extracted words is represented by \( F_i \) and the number of extracted words represented by \( n \). Table 3 shows some statistics of the word coverage rate of the two majors, with the result ranked according to the number of occurrence. As can be seen from Table 3, in computer science major, 10 words with the highest frequency have the coverage rate of 0.33217; in machine tooling major, 10 words with the highest frequency have the coverage rate of 0.35102. As the 10 words are also most frequently used in English, despite their small number, their coverage rate is very high when the number of word forms reaches about 200.

Table 3 Statistical Result of Word Coverage Rate

<table>
<thead>
<tr>
<th>Computer science</th>
<th>Machine tooling</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of word forms</td>
<td>Coverage rate</td>
</tr>
<tr>
<td>10</td>
<td>0.33217</td>
</tr>
<tr>
<td>40</td>
<td>0.41270</td>
</tr>
<tr>
<td>81</td>
<td>0.52148</td>
</tr>
<tr>
<td>126</td>
<td>0.56772</td>
</tr>
<tr>
<td>207</td>
<td>0.69381</td>
</tr>
<tr>
<td>567</td>
<td>0.72308</td>
</tr>
</tbody>
</table>

4. CONCLUSION

In language teaching, corpus construction and application are of great research value and practical significance. AI-based application of corpus can effectively improve foreign language teaching ideas and methods. Besides, data mining can be optimized by relying on common productive and framework knowledge representation methods and on certain algorithms. This can promote development of both the students and the teachers. AI technology can change the teaching model and teaching methods, and promote innovation of teaching ideas. Owning to its continuous improvement and development, AI technology can see more extensive application in
English teaching, not only in practical vocabulary teaching, but also in language research and teaching exploration.

REFERENCES