Design of an Intelligent Support System for English Writing Based on Rule Matching and Probability Statistics

https://doi.org/10.3991/ijet.v13i11.9608

Sa Wang, Hui Xu(✉)
Beihua University, Jilin, China
xuhui28193@163.com

Abstract—In view of the lack of intelligent guidance in online teaching of English composition, this paper proposes an intelligent support system for English writing based on B/S mode. On the basis of vocabulary, grammar rules and other corpus, this system uses Natural Language Processing technology, which combines rule matching and probability statistics, to evaluate and optimize the efficiency of the composition. The empirical results show that the system can effectively improve the teaching direction according to the results of intelligent quantitative analysis.

Keywords—Intelligent Support System; English Writing

1 Introduction

With the economic globalization and the rapid development of science and technology, international cultural exchanges are becoming increasingly frequent, so English, as an international language and communication tool, becomes increasingly important. For colleges and universities, it is necessary to train a new generation of students with higher English proficiency and comprehensive English application ability. This puts forward higher requirements for English teaching in higher education, which requires teachers to not only pay attention to the cultivation of students' innovative spirit, but also strengthen the cultivation of their comprehensive English proficiency. English writing ability is an important index to measure college students' comprehensive English application ability. However, one of the main reasons for students' low achievement in English exams is that they make more mistakes in vocabulary and grammar, and their scores are lower. This reflects that English writing has always been a weak link for English learners. With the continuous development of computer networks, computers and networks have a great influence on English teaching. Computer and network, as a supplementary teaching method, embodies many unique advantages. Making full use of the advanced Web information technology, many colleges and universities have begun to focus on the construction of Internet-based online education intelligent system to achieve the English teaching mode combining computer and classroom. As a result, it not only breaks the limitations of traditional education examination, but also students are not bound by time and place for writing training and evaluation, which brings great convenience to teachers and students [1]. Intelligent English writing system is used by college students as an assistant tool in the process of English writing, which can provide a perfect learning environment for learners and effectively improve their English writing level. Therefore, aiming at the needs of students' English writing, natural language processing, rule matching and probability statistics are applied to design an intelligent English writing training system, aiming at exploring a new method of efficient English writing training based on B/S, so as to improve the teaching effect of English writing and students' English writing level.

2 Literature review

With the rapid development of modern information technology and the acceptance of the concept of autonomous learning by more and more people, online learning has gradually become a novel way of English learning. The assistance of network technology can provide better conditions and environment for English learning. At the same time, students can be exposed to more professional and authentic English language materials, making the language learning process more effective [2]. At present, there are some maturely applied systems abroad, and different systems have different concerns. They adopt different algorithms and evaluation methods and each has its advantages and disadvantages, but the reliability and effectiveness of their practical applications are good [3].

Project Essay Grader (PEG) is the first English composition intelligent marking system in the world. PEG is the earliest intelligent teaching system used for English composition, which opens up a new world for English teaching, and its significance is naturally quite essential [4]. However, because the technology was not mature enough at that time, its operation efficiency was very low so that it could only reluctantly complete the grading work. This is because limited to the technical level at that time, PEG extracts few text features, mostly surface. Forsati and Shamsfard (2015) proposed that the learning activities of e-learning system have the characteristics of diversification. In order to solve the problem that students cannot make appropriate choices according to their personal conditions in the learning process, a fuzzy tree structure learning activity model was proposed. Based on the proposed model, the e-learning recommendation system prototype was carefully designed and developed. The good accuracy of the proposed method was proved, and the effectiveness of personalized e-learning recommendation system based on fuzzy tree matching in practice was also proved [5]. Roscoe et al. (2014), through analyzing that analytical language learning management system (LLMS) consists of three parts, namely, remote courseware, learned customer service system and data synchronization mechanism. The system supported English online learning from the perspectives of students' learning ability assessment, adaptive learning content, dynamic adjustment of learning strategies and feasible application of intelligent training system. LLMS effectively reflected students' individual differences in cognitive state and learning style, and further improved students' interest in English learning [6]. Ferrer et al. (2015) proposed a system used to detect lexical overlap pronunciation in English words spoken by English learners. The system was used by L1-English children to receive discourse training and to test English speeches made by L1-English children and L1-Japanese children with different English proficiency levels. A large number of test data show that L1 English pronunciation error rate is about 11%, L1-Japanese children speak English pronunciation error rate is about 20% [7]. Altani et al. (2017) pointed out that, in order to promote the scientific cognition of human learning and artificial intelligence, an intelligent agent was established to simulate human teaching and scientific learning. Based on this, an effective algorithm was proposed to obtain knowledge representation in the form of "depth features". This algorithm was integrated into machine learning agent SimStudent. The research showed that learning "depth features" reduced the requirement of knowledge engineering, and could be well applied to improve the teaching mode of tutoring system [8].

To sum up, the exploration of teaching assistant system under network environment is an important branch of educational informatization technology and system development. However, the related research content generally concentrates on the analysis of online learning mode and teaching situational strategy under network teaching environment. And the network intelligent learning system with more comprehensive functions lacks in-depth research. Therefore, based on rule matching and probabilistic statistics techniques, an intelligent English writing training system is designed. The use of intelligent English writing training system can effectively promote the overall reform of traditional teaching structure, educational model and even organizational structure, and achieve a completely new and efficient English writing training teaching mode.

3 Method and technology

3.1 Entity Framework

The intelligent English writing system designed here is based on B/S mode. The system is based on the Entity Framework as the basic development platform, which contains the content shown in Figure 1. Entity Data Model (EDM) is a mapping model between classes and databases. It contains three mapping files, namely concept model, drawing and storage mode. The Object Services layer converts the query into a command tree and passes it to Entity Client when implementing the query; it also converts the tabular data of the objects captured by the Entity Client layer to objects when returning the results; and it is also responsible for managing the state of the objects and tracking the changes of the objects. Entity Client mainly converts the queries of LINQ to Entities and Entity SQL into SQL statements. At the same time, it converts the database tabular data into object tabular data, and passes them to Object Services layer. There are many data tables involved in the design of the system, and applying the traditional database management will affect the development efficiency. Developers can use the Entity Framework to adequately define entities that map to database tables and use this entity directly in the business logic layer without writing some similar code, and the entity model (included in the EDM) can be modified and validated at runtime. Developers generally only need to operate the entity model, the framework will automatically complete the operation of the database, and the number of direct database management will be greatly reduced. Therefore, the use of Entity Framework in large-scale application systems can greatly reduce the development and maintenance costs of the project team.

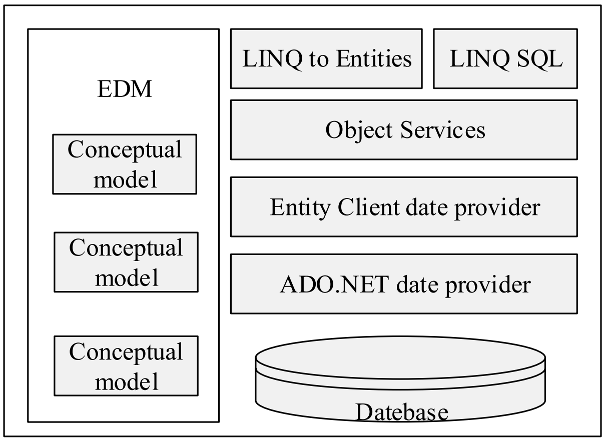


Fig. 1. Entity Framework overall framework

3.2 Natural language processing technology

Natural language processing, also known as computational linguistics, is a technology for processing natural language information. Natural language processing is an interdisciplinary field that includes computer science, artificial intelligence and linguistics. It studies all kinds of theories and methods that enable computers to process or "understand" natural languages. The purpose of natural language processing is to establish a variety of natural language processing systems, including machine translation system, natural language understanding system, automatic retrieval system, automatic text recognition system, database system and so on. Since the birth of the computer, human and computer interaction can only be achieved by programming language code, such as the use of basic, C, lisp and other computer programming languages. For a computer, it can only make different behavior responses according to binary instructions, and programmers often play a role of translation in the process. In other words, the functional requirements expressed in natural language are expressed in programming language, and then translated from a specific compilation into binary instructions that machines can understand. The natural language processing flowchart is shown in Figure 2.

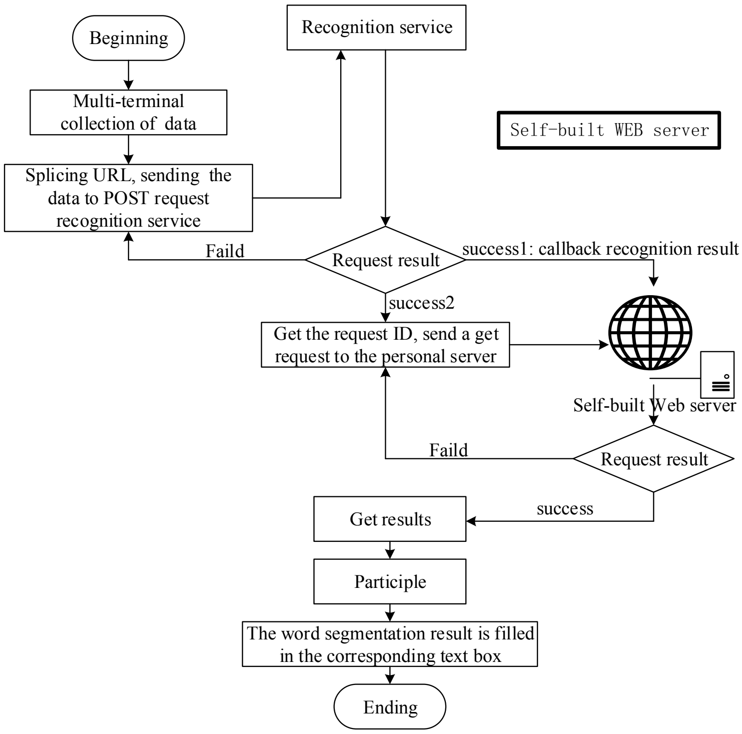

Beginning
Recognition service
Multi-terminal collection of data
Self-built WEB server
Splicing URL, sending the data to POST request recognition service
Faild
Request result
success1: callback recognition result
success2
Get the request ID, send a get request to the personal server
Self-built Web server
Faild
Request result
success
Get results
Participle
The word segmentation result is filled in the corresponding text box
Ending

Fig. 2. Natural language processing technology flow chart

Most grammar checking techniques in English writing rely on other basic techniques of natural language processing, such as clause, word segmentation, part-of-speech tagging system, and corpus technology. Sentence grammatical errors refer to the errors in writing for a single sentence. Writing errors that need to be context-related often involve semantic analysis, so "clause" processing is an indispensable pre-processing step before grammatical checking, and sentence segmentation accuracy has a certain impact on the final check accuracy. Word segmentation is the smallest and independent language unit in natural language. After the text is conducted with clause processing, there is an equally important pretreatment step called "word segmentation", which needs to use word segmentation technology. The main task of the technology is to separate each component of each sentence, including words and symbols, which play a very important role in it. Part-of-Speech tagging technology is to assign a correct part-of-speech tagging to a sentence word by word according to the information in the sentence context. It is also one of the key basic techniques in natural language processing. It is applied to various fields, including text indexing, text classification, corpus processing, and language synthesis. Corpus is a large-scale electronic text database scientifically sampled and processed. It consists of text data itself, text metadata and some linguistic annotations. Corpus is the main resource of empirical linguistic research methods, which is analyzed and processed through computer tools. It can be used in dictionary compilation, language teaching and natural language processing.

3.3 Checking algorithm based on rule matching

In the rule-based grammar checking method, the basis of judging whether a sentence has grammatical errors is to match the word sequence or part-of-speech sequence in the text with the rule base. Therefore, it is the key step of this algorithm to generalize and model rules and build rules library reasonably. The rule-based grammar checking algorithm is improved on the instance-based grammar checking algorithm, so the rule base is constructed by rules that generalize the wrong grammars. In the process of implementation, the grammar checking algorithm based on error rules matches the text to be checked with the error rule base. If the sentence in the text conforms to the error rule in the library, then the grammar error of the corresponding rule occurs in the sentence; otherwise, if the whole rule base cannot be matched successfully, then it shows that the sentence is correct.

3.4 Statistical probability checking algorithm

In the grammar checking algorithms based on statistics, the most widely used is the algorithm based on the n-gram model. N-gram grammar checking system is usually divided into two stages: training phase and checking stage. Firstly, a database with relevant N-gram grammar information is built, which is necessary for N-gram grammar checking. In order to successfully build a database, training corpus is usually necessary. After training the corpus, some N-gram information can be obtained, and then the information is stored into our database. Then, when judging whether the text is wrong or not, the information in the database just created is compared. All the probabilities of binary grammar that may appear in the corpus need to be accurately trained in advance and then added to the database. For the text checked, all binary grammar probabilities need to be extracted, and then the whole sentence probabilities, as shown in Formula (1), are equal to the multiplication of all binary grammar probabilities:

$$\mathrm{P(S) = P(< BOS >| \ word1\ ) * P(\ word1 \ | \ word2\ ) * P(\ word2 \ | \ word3\ ) *} \\ \mathrm{P(\ word3 \ |< EOS >\ )} \tag{1}$$

Before this, the system will pre-set a threshold, and the probability of the whole sentence and the system threshold are obtained and compared. If the probability P of the sentence is less than the threshold set by the system, the sentence can be judged to be wrong. If the probability P of the sentence is greater than the threshold set by the system, then the sentence can be judged to be correct.

3.5 Introduction to and working principle of Language Tool

Language Tool is a new, open-source, and extensible grammar correction system that combines English, French, German and other languages. Many researchers have studied and improved its English grammar proofreading XML rule base. The source code has been published on the Language Tool website. It can be downloaded or redeveloped free of charge according to the LGPL open source protocol. It focuses on the detection of complex English word morphological errors, word usage errors and syntactic errors, and provides corresponding error results. Language Tool adopts a separate and assembled system architecture, which integrates sentence segmentation, part-of-speech tagging, phrase chunking, word spelling checking, grammar checking and other functional modules. It has the characteristics of intuitive structure, high freedom and high openness.

Language Tool requires the design of independent grammatical error rule files based on XML markup language and conforming to its format requirements for various natural languages. It mainly uses regular expressions for pattern matching, word segment matching, part-of-speech combination matching and even hybrid matching to describe the grammatical errors. Some special grammars that are difficult to describe in a simple XML format also support the use of programming languages to write special grammar files for matching judgments. The grammar description system defined by Language Tool makes it relatively easy to convert the complex grammar of natural language into a description format that can be recognized by the system. The software can read and use directly it without modifying the program source code. Although the rule-based approach mentioned above is unlikely to exhaust all grammatical phenomena, the Language Tool's ability to perform spell proofreading and grammar checking is quite good as an open source software, coupled with researchers' constant updating of the rule base through simple rule editing tools.

When the program starts to work, the main program will first call the English grammar rules library file, initialize the system environment, and wait for input text. If the English text to be detected is a paragraph, the clause module is called first and a separator is added at the beginning and end of each sentence. Then, the resulting sentence is stored in the list, waiting for word segmentation, part-of-speech tagging, disambiguation, and phrase chunking. If it is not a paragraph, it will work directly on the sentence. After word processing is completed, the tagged prefix is matched with the first rule in the rule base. If the match is successful, the position of the prefix is wrong; if the match is unsuccessful, the next word is selected and matched with the first rule. It is cycled and repeated until all the words in the sentence match with the first rule. Then, save the word that is not matched successfully to the temporary list and match it with the next rule. Determine whether all the rules match, and if not, it is necessary to determine whether the part-of-speech matches successfully. It is complete until all rules are matched. Finally, all the analysis results of successful matching, including all possible error locations, error descriptions and corresponding modification suggestions, form a list of error feedback to the user.

4 Results

4.1 System requirement analysis

With the continuous development of computer network and network technology, the original mode of writing and testing training has been unable to meet the needs of large-scale students' online English learning and testing. In addition, students' English test assignments will be no longer tested or corrected in the traditional paper-based way, but use the online learning and testing with rapid communication. It becomes very important for students to need more English writing methods and testing resources.

According to the requirement of cultivating students' comprehensive English application ability in colleges and universities, an online English learning and testing system is constructed. It not only has the common functions of the general learning platform, but also needs the characteristics of an intelligent testing system.it mainly includes: firstly, in the system, students can test after the learning is finished, and after the test is completed, the test training is submitted. Then, the system which belongs to the objective question gives the test results immediately and points out the wrong points and the knowledge points covered. The subjective writing test is that, after the students submit the test results, teachers assign the time to review the papers and give online corrections or reply by e-mail. In the analysis of English test, it is necessary to analyze the students' writing in the aspects of grammar, diction and situation, so as to help them find out the common mistakes and loopholes in English writing and the defects and misunderstandings in English learning, so as to improve the students' learning level and stimulate their learning motivation. Secondly, in the test training stage, the test questions in the question bank should be classified according to the difficulty program coefficients of the application ability level. When students pick up the test questions in the process of testing on the Internet, they can generate the test questions randomly. The students can make system self-selectin and capacity division test for the English application writing level according to their own learning English situation.

At present, most college students use English writing and testing intelligent systems more centralized, so the system needs to be able to ensure a normal and stable running state in the case of a large number of students simultaneously online learning and testing. For the system response time, test results timely storage and other system performance requirements are strict and data consistency is needed.

4.2 Overall framework design of intelligent English writing system

According to the choice of the system development platform, the frame design of the intelligent English writing system is shown in Figure 3. It is mainly divided into four levels for design.

The first level: data layer. This layer is mainly composed of English writing training database and English writing related documents. The English writing training database also includes the system-related writing test questions database and writing resource pool. Item bank and resource pool are linked by writing knowledge points. This system not only provides a data storage scheme, but also provides a file storage scheme. It stores some resources related to writing training but unable be input into the database in the form of files, and accesses the file path in the database.

The second level: the physical layer. Using Entity Framework, developers can fully define the entities mapped to database tables, and use this entity directly in the business logic layer. Generally, developers only need to operate the entity model, and the framework will automatically complete the operation of the database, greatly reducing the number of direct database management.

The third level: business logic layer. At this level, the classes, attributes and methods of the relevant business logic should be defined to complete the core functions. According to the analysis of the whole function module, it can be divided into four modules: writing assignment, writing test, writing skill training, and writing feedback. It mainly contains the businesses related to writing questions and writing resources and related functions of announcement management. Writing skills training includes skills training, writing assistance, resource recommendation and error recording business. This is where clauses, participles, spelling checks, and syntactic parsing logic processing involved in automatic corrections in writing feedback are completed.

The fourth level: presentation layer. According to the functional modules, the corresponding display pages and necessary interaction functions are developed. It mainly includes user pages of the above five functional modules. At this layer, user interface processing technology is used to improve user experience comfort and interaction. Especially when in writing feedback, teacher manual annotation problem, the user interface processing technology can be used in this layer to solve the problem.

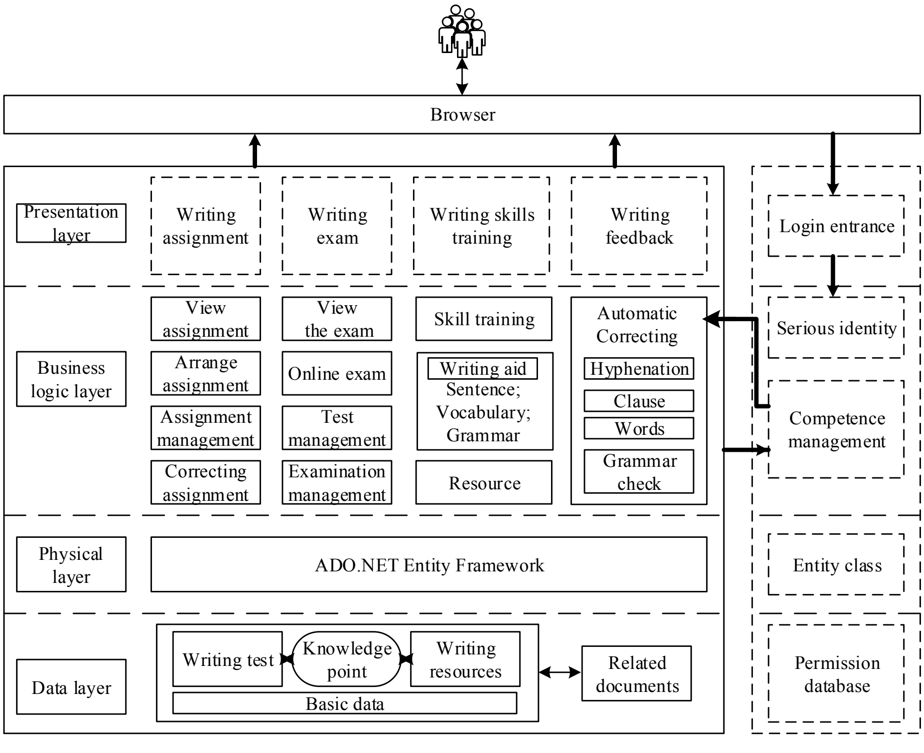


Fig. 3. The overall structure of the intelligent English writing training system

4.3 Knowledge base design

By constructing a knowledge base of English writing, the knowledge point in English learning is like a node on the network. The learning process or the explaining process of a knowledge point can be extended. Therefore, these knowledge points are used to mark the English writing test questions and writing resources, so as to establish a link between them and to help students smoothly start writing learning. The test questions and knowledge points can be linked by attributes, as shown in Figure 4. Question banks are associated with writing skills through skill attributes, keyword attributes are associated with vocabulary, sentence pattern attributes are associated with sentence patterns, and grammatical point attributes are associated with grammar.

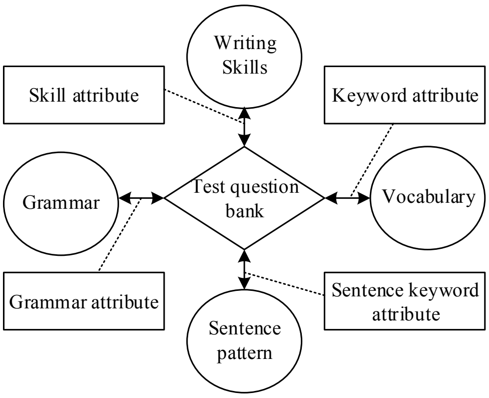


Fig. 4. The relationship between knowledge points and test questions

4.4 Implementation of skills training in intelligent English writing system

Students are trained independently by writing skill training module, and the main interface contains the function of system design. Writing skills, writing training, and writing materials are the main functions of this module, and the close relationship among them is described as shown in Figure 5. Writing skills use writing materials when explaining knowledge points and tell users how to use them; in writing training, it is necessary to master certain writing skills, because writing skills tell users how to train; writing materials provide the necessary resources for writing training, and at the same time, writing training can use the corresponding materials.

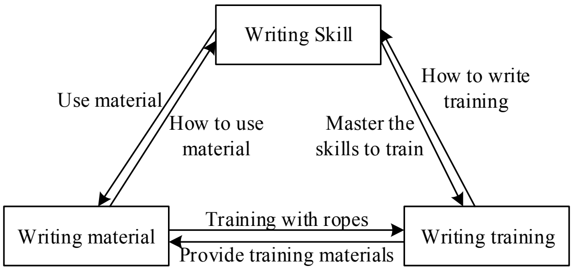


Fig. 5. The relations between writing skills, writing materials and writing training

When students enter the writing training interface, the system time starts. At the same time, it is possible to read the writing guidance and check the writing skills and model essay. In the training process, the structure of the article can be completed through the auxiliary function of writing.

4.5 Realization of automatic modification of writing in intelligent English writing system

This system calls the function of writing automatic correction in many places. Taking the automatic correction of writing training as an example, the specific process of implementation is illustrated. Students write online, and submit to the system for correction after the completion of the writing. The system calls Language Tool to achieve lexical and grammatical checking. The error messages returned by the Language Tool are strings, and for convenience of invocation, a data formatting model is especially defined in the entity class to store and format the information returned by the Language Tool. When processing the text, the information is taken from the model, plus the error and the suggested markup and the specific content. Finally, the presentation layer completes the work of page display and style rendering. The process described above is shown in Figure 6.

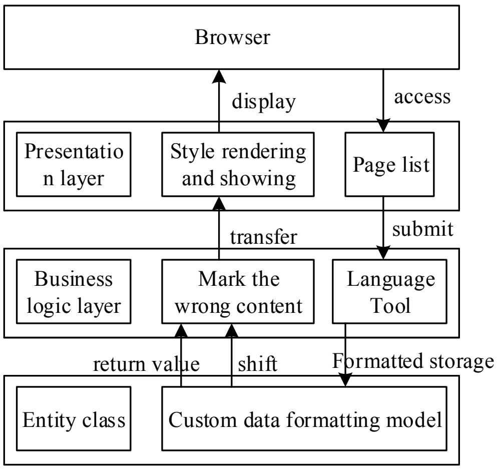
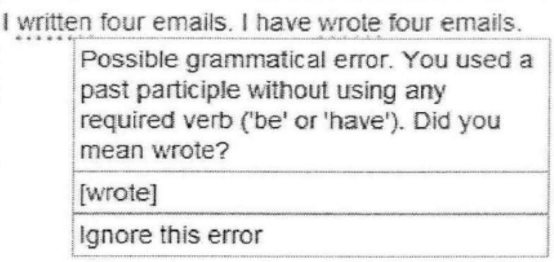


Fig. 6. English writing automatic correction process

As shown in Figure 7, after students use the automatic correction, the system will mark the error area with dotted lines. Clicking on the area, students can also see the error details and modification suggestions. Users can choose "Ignore this error" and the dotted line will disappear.



I written four emails. I have wrote four emails.

Possible grammatical error. You used a past participle without using any required verb ('be' or 'have'). Did you mean wrote?

[wrote]

Ignore this error

Fig. 7. Error tips and suggestions for changes

5 Conclusion

With the deepening of international cultural exchanges, many colleges and universities and educational institutions attach great importance to the comprehensive English application ability of the new generation of students. Advanced science and technology are applied to English writing to achieve a new type of training reform in intelligent English writing, which is transformed from the original single teaching mode to the integrated learning mode using information technology and network technology. Combined with the Entity Framework, the natural language processing technology is used to integrate the rule matching technology and statistical probability into the intelligent English writing training system based on B/S mode. By analyzing the requirements of the current intelligent English writing training system in colleges and universities, an effective set of intelligent English writing is designed. The four-layer structure of the system is further analyzed and a knowledge base of English writing is built. Through the implementation of the system, it is concluded that the intelligent English writing training system can effectively combine English writing skills, writing training and writing materials, and realize the function of automatic correction in error correction. This improves students' English learning ability and writing ability to a large extent.

6 Acknowledgment

You may mention here granted financial support or acknowledge the help you got from others during your research work.

7 References

[1] Stratos, K., Collins, M., & Hsu, D. (2016). Unsupervised part-of-speech tagging with anchor hidden markov models. Transactions of the Association for Computational Linguistics, 4: 245-257.

[2] Sidorov, G. (2013). Syntactic dependency based n-grams in rule based automatic Eng-lish as second language grammar correction. International Journal of Computational Linguistics and Applications, 4(2): 169-188.

[3] Nguyen, D. Q., Nguyen, D. Q., Pham, D. D., & Pham, S. B. (2016). A robust transformation-based learning approach using ripple down rules for part-of-speech tagging. AI Communications, 29(3): 409-422. https://doi.org/10.3233/AIC-150698

[4] Bhowmik, T., Niu, N., Savolainen, J., & Mahmoud, A. (2015). Leveraging topic model-ing and part-of-speech tagging to support combinational creativity in requirements engineering. Requirements Engineering, 20(3): 253-280. https://doi.org/10.1007/s00766-015-0226-2

[5] Forsati, R., & Shamsfard, M. (2015). Novel harmony search-based algorithms for part-of-speech tagging. Knowledge and Information Systems, 42(3): 709-736. https://doi.org/10.1007/s10115-013-0719-6

[6] Roscoe, R. D., Allen, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (2014). The Writing Pal intelligent tutoring system: Usability testing and development. Com-puters and Composition, 34: 39-59. https://doi.org/10.1016/j.compcom.2014.09.002

[7] Ferrer, L., Bratt, H., Richey, C., Franco, H., Abrash, V., & Precoda, K. (2015). Classification of lexical stress using spectral and prosodic features for computer-assisted lan-guage learning systems. Speech Communication, 69: 31-45. https://doi.org/10.1016/j.specom.2015.02.002

[8] Altani, A., Georgiou, G. K., Deng, C., Cho, J. R., Katopodi, K., Wei, W., & Protopapas, A. (2017). Is processing of symbols and words influenced by writing system? Evi-dence from Chinese, Korean, English, and Greek. Journal of experimental child psychology, 164: 117-135. https://doi.org/10.1016/j.jecp.2017.07.006

8 Authors

Sa Wang and **Hui Xu** are with Beihua University, School of Foreign Language, Jilin, China (xuhui28193@163.com).

Article submitted 23 September 2018. Resubmitted 17 October 2018. Final acceptance 26 October 2018. Final version published as submitted by the authors.