Design and Implementation of Automatic Examination Scoring System Based on Natural Language Processing

Wenyan Yang(✉)

Dalian Vocational and Technical College (Dalian Open University), Dalian, Liaoning, China
657942541@qq.com

Abstract. Based on natural language processing and the development method of Python Web in the Flask framework, this paper completes the design and implementation of the automatic examination scoring system. The key algorithms such as lexical analysis, syntactic dependency analysis, and knowledge extraction under the system natural language processing technology are cited to complete the similarity analysis of keywords and sentences in Chinese text content, so as to realize the automatic scoring function of subjective answers in online examination content. Under this system, teachers can greatly simplify the work flow of manual marking with the help of "scoring standard setting" and "automatic scoring" functions, avoid the influence of human errors and uncertain factors in the process of manual marking. Practically improve the completion quality and comprehensive management level of education and teaching tasks. At the same time, the system also makes a new attempt for the rapid development of online education and teaching and the reform of teaching methods.

Keywords: natural language processing · automatic scoring system · similarity analysis · Python Web

1 Introduction

With the rapid development of network information technology and the increasing popularity of mobile terminal devices, digitalization, informationization and intelligence have become the new trends of development in all walks of life in the economy and society. In the field of education industry, the innovative integration of network information technology gave birth to a new type of education model-online teaching. The rise of online teaching marks the beginning of deepening the reform of modern education in China, and also promotes the comprehensive change of educational system and mechanism in the new era. Since the outbreak of the epidemic in 2020, online teaching has become the main front of education and teaching activities. With its advantages of convenience, high efficiency, abundant resources and various forms, it not only meets the requirements of education and teaching under special circumstances, but also coexists with offline classroom teaching mode for a long time, becoming an effective supplement to classroom teaching. From the analysis of the development trend of the whole online

Z. Zhan et al. (Eds.): ICBDIE 2022, AHCS 5, pp. 1048–1058, 2023.
https://doi.org/10.2991/978-94-6463-034-3_107

teaching, the development form of online teaching has realized the transformation from the third-party live video software to the specialized online teaching platform. Under the specialized platform, the functions cover online teaching, homework release, questions and answers, communication and feedback, and also support online testing. The appearance of online test function meets the needs of periodic study results inspection in the teaching process. Paperless test method saves a lot of time in the links of test questions, printing, organization, marking and scoring in traditional tests, and greatly improves the efficiency and quality of education and teaching [9].

However, based on the analysis of the actual application process of online testing function under the current online teaching platform, there are still some problems that restrict the improvement of the efficiency of online testing function. At present, the scoring technology of objective questions in online testing system is mature, and students can complete the scoring by comparing the input answer data with the preset standard answer data in the database. However, the evaluation of subjective questions mainly depends on manual evaluation, which takes up teachers' time and increases teachers' workload. In view of this, this paper holds that the core technology based on natural language processing can be used to identify and subdivide students' answers, extract words, sentence and keywords, and calculate the similarity between sentences, keywords and standard answers of students' answers by using various algorithms, so as to obtain corresponding weighted scores and realize the automatic scoring function of subjective questions in exams. In addition, based on Python Web development technology, the test automatic scoring system is constructed, which is convenient for teachers to complete the system login and function call from the web client browser. On the Web server side, it will be designed with the Flask framework, and natural language processing (NLP) function module will be called through API data interface to realize the specific functions of the system. The construction of automatic scoring system can not only reduce the teachers' heavy marking work, but also control the interference of external human factors, and basically realize the fairness and fairness of marking. [8] It also improves the systematic adaptability of online teaching platform for holding large-scale student examinations.

2 Introduction of Key Technologies

2.1 Natural Language Processing

As a unique characteristic of human beings, language is not only an expression of communication, but also a carrier of the spread and development of the achievements of human civilization. For this reason, the study of natural language has never stopped. Natural Language Processing (NLP) is an important direction in the field of computer science and artificial intelligence. It studies various theories and methods that can realize effective communication between people and computers in natural language. [7] Natural language processing is a scientific field integrating linguistics, computer science and mathematics. Its main purpose is to develop a computer system that can effectively realize natural language communication. Under the system, in order to enable people to use and control computers in their own language, we must first enable computers to understand the meaning of natural language; Secondly, computers should be able to express corresponding intentions and thoughts in natural language. Therefore, natural

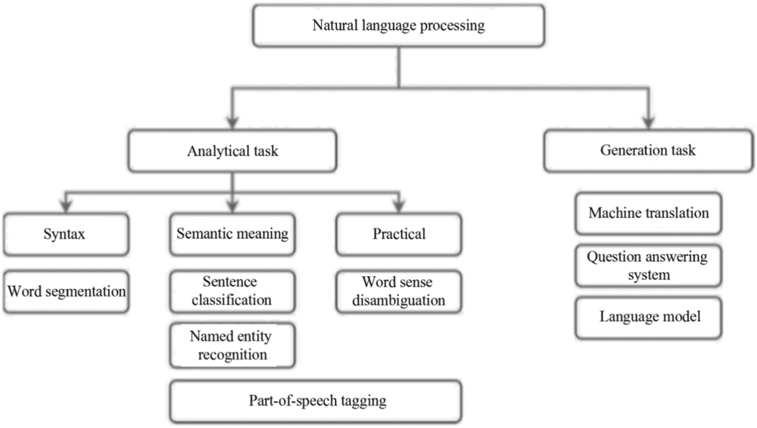


Fig. 1. Natural language processing application type

language processing can be divided into two parts: natural language understanding and natural language generation. The operation of the whole process needs to use computer system and software technology to complete the analysis, calculation and statistical operation of natural language from morpheme, pronunciation, corpus, grammar, semantics, pragmatics and other dimensions, and take this theory as the core to expand the application of NLP into many types. As shown in Fig. 1, the natural language understanding part includes word segmentation, word sense disambiguation, named entity recognition, part-of-speech tagging, sentence classification and so on, and the natural language production part includes language model generation, question answering system, machine translation and so on [10].

For the automatic examination scoring system described in this paper, it is mainly aimed at the automatic analysis and processing of subjective answers and the calculation of scoring scores in the online test function under the network online teaching platform, which belongs to the research category of natural language understanding. The overall language environment is Chinese, and its main processes are as follows.

2.1.1 Sentence Segmentation Processing

The core task of sentence segmentation processing is to divide the input text content into several clauses according to the established marks. Many marks are punctuation marks that can indicate the completion of sentence meaning, such as period, question mark, exclamation point, semicolon, etc. Sentence segmentation processing can transform the basic language units expressing complete meaning from paragraphs into sentences, and with less text content and simpler logical relations, it can make the processing of text information simpler and more effective, so as to improve the system's grasp speed of the meaning of the whole text.

2.1.2 Analysis of Words and Sentences

After sentence segmentation processing, the obtained clauses are segmented according to the usage standard of Chinese vocabulary, and multiple independent words are obtained, which is called Chinese word segmentation. For example, "I love you China!" Getting Chinese word segmentation of "I", "Love", "You" and "China" after sentence segmentation is the premise to further refine the text content and promote subsequent analysis and processing. For the complexity of the application of Chinese characters and words, Chinese word segmentation needs to solve two technical difficulties: ambiguous word segmentation and new word recognition. Among them, ambiguous word segmentation exists in many situations, including combination ambiguity, intersection ambiguity, true ambiguity and so on. In addition, for the recognition of new words, it is more important to study the acceptance and application of new words, such as the annual popular online words "involution" and the annual news hot words "breaking defense" and so on, which need to be continuously incorporated into the dictionary in order to achieve accurate Chinese word segmentation. Common methods of word segmentation include dictionary matching and probability statistics of adjacent words. After the completion of Chinese word segmentation, part-of-speech tagging will be carried out, that is, nouns, verbs, adjectives, adverbs and other categories will be distinguished. Some proper names will also be identified independently, such as names of people and places. Finally, in the grammar analysis, through the analysis and identification of the subject, predicate, object and other components in the sentence, the collocation relationship between words and syntax is presented.

2.1.3 Keywords Extraction

After sentence segmentation, word segmentation and syntactic analysis, the relationship between all words in the sentence and the grammatical structure of the sentence is obtained. At this time, keyword extraction algorithm is used to obtain keywords in sentences, and function words and stop words in sentences are eliminated at the same time to reduce the influence on the accuracy of subsequent similarity calculation. Generally, keywords refer to nouns, verbs, adjectives, etc. in sentences, which can express meaning as close as possible to the text and form a high degree of conciseness of the text information. In this paper, TF-IDF keyword extraction algorithm is adopted. Its core idea is that a word appears frequently in one article and rarely in other articles, so it is considered that the word can better represent the meaning of the current article. That is, the importance of a word is directly proportional to the number of times it appears in documents, and inversely proportional to the frequency of its occurrence in documents in corpus. [6] As shown in Formula (1), TF-IDF algorithm generates corresponding weighted values by weighting all candidate words in the text content, and arranges the weighted values in descending order, and the first few words with the highest weighted

Table 1. Comparison table of advantages and disadvantages of common sentence similarity algorithms

Method	Advantage	Disadvantage
Based on string matching	Simple principle and easy realization	Ignoring word meaning information
Based on large-scale corpus	Strong objectivity	Dependent corpus
Ontology-based knowledge	Easy to understand	Limited by dictionary, ignoring syntactic structure

values are the keywords of the text content.

$$\begin{aligned}
&\text{Word frequency (TF)} = \frac{\text{Number of times the word w appears in the document}}{\text{Total number of times in document}}\\
&\text{Inverse document frequency (IDF)} = \log\left(\frac{\text{Total number of documents in corpus}}{\text{Text with word W} + 1}\right)\\
&\text{TF} - \text{IDF} = \text{TF} * \text{IDF}
\end{aligned} \tag{1}$$

2.1.4 Similarity Calculation

After extracting keywords, it is necessary to complete the similarity between keywords and standard corpus vocabulary. In this system, the scores of the students' answers are determined by the similarity between the keywords in the students' answers and the keywords in the reference answers. The higher the similarity, the higher the score; otherwise, the lower the score. The commonly used methods of keyword similarity calculation include corpus-based calculation, search engine-based calculation and ontology-based knowledge word similarity calculation. Among them, the algorithm based on ontology knowledge words mainly calculates the semantic distance according to the hierarchical structure in the current core application dictionaries such as HowNet and Synonym Word Forest. The semantic distance is inversely proportional to the keyword similarity, and the keyword similarity can be calculated by calling the algorithm [1].

In addition, in order to improve the accuracy of the automatic scoring system, the system also adds sentence similarity calculation. The system supports the comparative calculation of similarity between the semantics expressed by one or more sentences in the clause result and the sentences in the topic reference answer after the students answer the text content clause operation, and obtains the similarity value according to the corresponding algorithm. As shown in Table 1, in order to compare the advantages and disadvantages of sentence similarity algorithms commonly used in the current practical application, the system selects the calculation method based on string matching, combining the advantages and disadvantages of each method with the actual requirements of the system. Word2Vec model is adopted to take the repetition and co-occurrence of characters or words in different sentences as a measure of sentence similarity.

2.1.5 Scoring Formula

After the system keyword similarity calculation and sentence similarity calculation, the final scores of the students' subjective questions are obtained after the scores of the two parts are calculated by the scoring formula. The setting of the scoring formula will require that the error be reduced as much as possible, and the algorithm of weighted proportion will be adopted and the error parameter value F will be introduced to calculate the final score. The calculation formula is shown in formula (2). In the formula, g stands for the final score, p is the total score of the current topic, k is the keyword similarity Kp is the weight of the keyword score, s is the similarity of the X sentence, and Sp is the weight of the X sentence.

$$G = \left(K * K_P + \sum\nolimits_{\mathrm{x}=1}^{\mathrm{n}} S * S_P\right) * P + F \tag{2}$$

2.2 Python Web Technology and Flask Framework

Python is a high-level scripting language that combines interpretability, compilation, interactivity and object-oriented. Since its release, it has become a programming language that can be applied to multiple development platforms by virtue of its minimalist design concept and standardized interaction mode, and has gradually completed the updating of various functions to realize the development of various types of independent large-scale projects. For Web application development, Python has many mature Web development template technologies. The application of Django, Flask, CherryPy, Pyramid and other development frameworks can enable developers to develop system functions with less code, which not only greatly improves the efficiency, but also speeds up the running speed of the system.

Flask framework is a lightweight Web framework, which is more flexible and extensible than other frameworks. The core of Flask framework is Werkzeug WSGI toolkit and Jinja2 template engine. The operation process is shown in Fig. 2. After the user sends an HTTP request to the Server through the client browser, WSGI Server forwards the request to WSGI App, which completes the logical processing of the request and returns the processing result to the server. WSGI App can include several stack middlewares, which need to implement both Server and App at the same time, so they can play a regulatory role between Wsgi server and wsgi application: for server, middleware acts as application program, and for application program, middleware acts as server [4].

2.3 Development Environment

According to the usage requirements of the above related application technologies, complete the construction and deployment of the development environment. The system is designed and developed by Python Web technology, and the operating system is Windows10.0 and the language development environment Python 3.6. Under the Windows system, the third-party tool Virtualenv is used to create a virtual environment, and the installation and deployment of the Flask framework is completed, Flask version 2.0. Under PyCharm, open the virtual environment Python.exe created by Virtualenv to

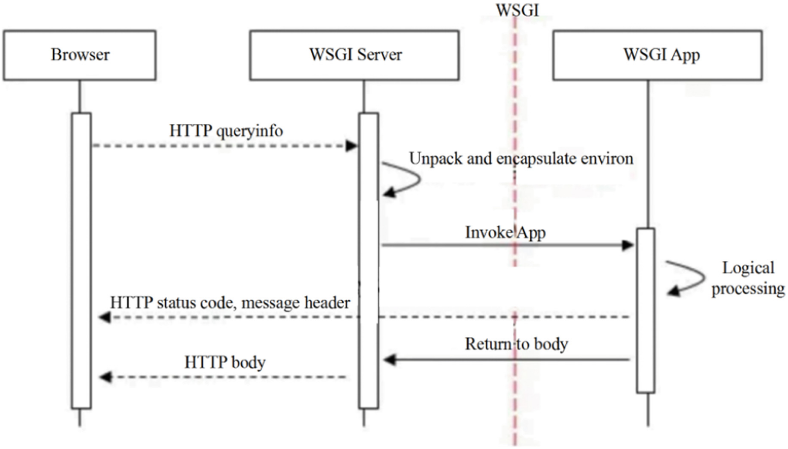
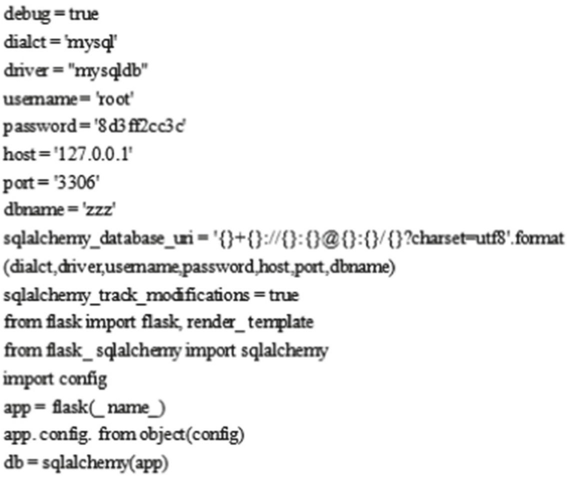


Fig. 2. Flask operation flow chart



```
debug = true
dialct = 'mysql'
driver = "mysqldb"
usemame= 'root'
password = '8d3ff2cc3c'
host = '127.0.0.1'
port = '3306'
dbname = 'zzz'
sqlalchemy_database_uri = '{}+{}://{}:{}@{}:{}/{}?charset=utf8'.format
(dialct,driver,usemame,password,host,port,dbname)
sqlalchemy_track_modifications = true
from flask import flask, render_ template
from flask_ sqlalchemy import sqlalchemy
import config
app = flask(_name_)
app. config. from object(config)
db = sqlalchemy(app)
```

Fig. 3. Key code of Flask framework connection database

complete the construction of the Flask project. The system server selects MySQL, and completes relevant configuration by declaring the database user name, password, host address, etc., and realizes various operations on MySQL data through Python-MySQLdb. The key code is shown in Fig. 3. The web server of Python consists of uWSGI server and Nginx server. Among them, Nginx server is responsible for user demand distribution and overall balanced load control of the system; The uWSGI server supports the communication between WSGI Server and WSGI App under the framework of Flask. [3] Through the introduction of the above key technology theories, we have determined the overall environment of system development, the configuration of related software and tools, and also made clear the technical feasibility of the overall project of the automatic examination scoring system.

3 Requirements Analysis

3.1 System Requirements Analysis

The automatic examination scoring system based on natural language processing will solve the problem that subjective questions can't be scored automatically in the current online examination, and a series of operations such as sentence segmentation, word segmentation, keyword extraction, word similarity calculation, sentence similarity calculation and total score calculation will be carried out for students' Chinese answers. Through the construction of the system, the workload of teachers in the process of reading and evaluating can be greatly reduced, and the work efficiency can be improved.

The automatic examination scoring system supports teachers' users to log in and use their unique identity information after registering and authenticating with their accounts. According to the teacher's actual operation process in the process of reading and evaluating, the system functions are set, including three functions: scoring standard setting, automatic scoring and score viewing. Among them, under the scoring standard setting function, teachers can set the scoring related parameters and contents by themselves, so as to facilitate the use of the subsequent automatic scoring function module. Under the automatic scoring module, the system will automatically complete all operations under natural language processing, including the call of external services, the start and stop of several algorithms of the system, and output the final result. Under the score viewing function, teachers can view the comprehensive scores of students' answers, and support manual verification to reduce the error between automatic scoring and manual scoring.

3.2 Global Design

The automatic scoring system will adopt B/S architecture, based on Python Web, and adopt Flask framework to complete the overall design and development of the system according to MVC pattern. According to the business requirements of the system, the system is divided into three modules, namely, input module, calculation module and output module. The overall architecture diagram is shown in Fig. 4. In the input module, the input contents include the retrieval of students' answers, the setting of relevant parameters, the input of reference answers and the adjustment of error parameter value F. Under the calculation module, the system will call external services to complete sentence segmentation and word segmentation. Among them, the sentence segmentation and word segmentation operations use NLPIR Chinese word segmentation system of Chinese Academy of Sciences, and access is realized in the form of Web service interface, which can be directly called across platforms without downloading SDK. [5] Keyword extraction adopts TF-IDF algorithm, and keyword similarity is calculated by semantic similarity based on HowNet corpus. For syntactic analysis of sentences, the LTP service platform of Harbin Institute of Technology is adopted, and the service also supports the remote call of API data interface. Sentence similarity calculation relies

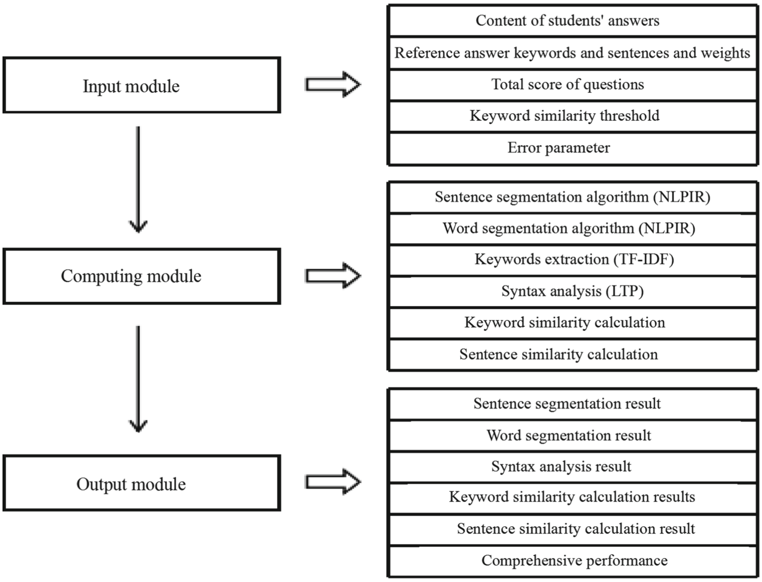

Input module
Content of students' answers
Reference answer keywords and sentences and weights
Total score of questions
Keyword similarity threshold
Error parameter
Computing module
Sentence segmentation algorithm (NLPIR)
Word segmentation algorithm (NLPIR)
Keywords extraction (TF-IDF)
Syntax analysis (LTP)
Keyword similarity calculation
Sentence similarity calculation
Output module
Sentence segmentation result
Word segmentation result
Syntax analysis result
Keyword similarity calculation results
Sentence similarity calculation result
Comprehensive performance

Fig. 4. System overall framework diagram

on Word2Vec model to calculate the similarity of words in sentences separately and to calculate the similarity of sentences by integrating the features of word dependency in syntactic analysis. After the calculation of the calculation module, in the output module, the system displays the results of sentence segmentation, word segmentation, syntactic analysis, keyword similarity calculation, sentence similarity calculation and comprehensive score calculation. Finally, after checking and verifying by teachers and users, all data are stored in the database to complete the data persistence operation.

4 Detailed Function Realization

4.1 Scoring Standard Setting

Under this function module, teachers and users can set different scoring standard parameters and input reference answer content, keywords or key points according to different topics. Among them, the standard parameters include the keyword weight Kp and sentence weight Sp in the reference answer, as well as the total score P and error parameter F of the title. In actual operation, multiple keywords or key points are separated by semicolons, and the setting of relevant weights should be controlled within a reasonable range of 0–1. [2] As far as this system is concerned, the weights of keywords and sentences are 0.5 to facilitate the subsequent calculation.

4.2 Automatic Scoring

Under this function, the teacher can select the screening and calling of students' answers in batches in the page, and perform batch automatic scoring operation with one click. The system will automatically calculate the scores of the preprocessed reference answers and examinee answer texts from two aspects: keyword score and sentence expression score. Among them, when calculating the keyword score, the system sets the distinguishing threshold of the keywords to 0.7. That is to ensure that the students answer content keywords and reference answer keywords have a more obvious distinction, and give a reasonable calculation score.

4.3 Achievement Inquiry

Under this function, teachers can intuitively see the analysis and statistics of students' answers after the automatic scoring operation. Teachers have the right to modify the final comprehensive scores of students, and can make appropriate adjustments to ensure the scientificity and rationality of the system results in case of large errors or special scores.

5 Conclusion

Based on natural language processing and the Python Web development method of the Flask framework, this paper completes the construction of the examination automatic scoring system. Through the test and application of this system, the key algorithms such as lexical analysis, syntactic dependency analysis and knowledge extraction can be called normally, which supports the system to automatically score subjective questions in online examinations. It not only effectively reduces teachers' daily workload, improves teachers' work efficiency, but also eliminates the bad factors in the examination score, which increases the scientificity and rationality of the score. The overall design of the system has a high expansion, which can be adjusted simply to adapt to the test scores of different subjects or majors. It can effectively improve the completion quality and comprehensive management level of education and teaching tasks, and also make a new attempt for the rapid development of online education and teaching and the reform of teaching methods.

References

1. Dong Yuan, Qian Liping. (2017) Text Similarity Calculation Based on Semantic Dictionary and Word Frequency Information. Computer Science. 11.
2. Guo Qingchun. (2017) Design and Implementation of English Composition Automatic Scoring System.Harbin: Harbin Institute of Technology. 12.
3. Li Chao et al. (2019) A Web Server-side Design Based on Python Flask. China Computer & Communication (Theoretical Edition). 04.
4. Niu Zuodong, Li Handong. (2019) Building a Practical MVC Framework that Can be Developed Efficiently Based on Python and Flask Tools. Computer Applications and Software. 07.
5. Qi Xiaoying. (2019)Semantic Intelligence Analysis of Artificial Intelligence News Events Based on NLPIR. China Computer & Communication (Theoretical Edition). 10.
6. Shi Fenggui. (2020) Implementation of Chinese Text Classification Based on TF-IDF. Modern Computer. 02.

7. Song Yifan. (2019) The Development History and Present Situation of Natural Language Processing. China's High-tech. 02.
8. Wang Xianze. (2018) Research on Technical Progress of Automatic Scoring System in Education. Educational Measurement and Evaluation. 05.
9. Wu Hao. (2020) Analysis of Online Examination System of Distance Education under Computer Network Environment. Digital Communication World. 04.
10. Xue Yafei. (2018) Deep Learning for Natural Language Processing.Electronic Technology & Software Engineering. 06.