Research Article

Modeling of the English Translation Complex System Based on Discrete Data Mining

Ruirui Bian (ORCID)

Henan Industry and Trade Vocational College, Zhengzhou, Henan 451191, China

Correspondence should be addressed to Ruirui Bian; 201309020216@stu.sdnu.edu.cn

This paper provides an in-depth analysis and research on the application of web discrete data mining in modeling complex systems for English translation. This paper adopts association rule mining, clustering, and classification methods to deeply analyze English translation, mine the interesting rules in it, find out the interinfluence relationship of system modeling and the relationship between translation behaviors, analyze and compare different data mining methods, and make an evaluation of the mining results to provide a scientific reference basis for the relevant decision-making scheme of the English translation. The paper elaborates the relevant concepts of data mining and the theoretical basis and analysis process of association rule mining, clustering rule mining, and classification rule mining methods. The application of data mining techniques in modeling and analysis of complex systems of the English translation is highlighted, and these three mining methods are used to analyze a set of English sample data: through the analysis of the correlation of different English translations, some feasible suggestions are provided for the order of translation settings; through the analysis of clustering of English translations, the shortcomings of traditional translation methods are made up; through the analysis of the factors affecting the accuracy of English translations factors, a useful reference is provided on how to improve English translation and then improve the quality of translation.

1. Introduction

With the development of information technology and database technology, data warehouses and databases have been widely used in different fields, such as product sales, information management, and enterprise management. The rapid growth of data volume has put forward new requirements for data query, data storage, and data management, and there is an urgent need for a new tool to intelligently mine valuable information and knowledge from the massive data. With the opening of translation majors in colleges and universities, while more and more English majors and nonforeign language majors are offering translation courses, translation teaching is gradually getting attention. However, there are still many problems in the cultivation of translation talents in colleges and universities in terms of teachers, teaching time, and teaching mode. With the increasing maturity of the field of information-based education and the increasing number of students, the school academic management information management system stores a large and rich teaching-related data, and the use of data mining technology to mine valuable information from this large amount of data can provide important reference information for the rationalization of teaching management and the scientific formulation of teaching syllabus in the future, which will help the reform and development of education.

With the opening of translation majors in colleges and universities, while more and more English majors and nonforeign language majors are offering translation courses, translation teaching is gradually gaining attention [1]. However, there are still many problems in the cultivation of translation talents in colleges and universities in terms of teachers, lecture time and teaching mode, and so on. Since the 21st century, the web 2.0 era has emerged, which has brought inspiration to the reform of English translation and provided unlimited possibilities for the benign and sustainable development of English translation [2]. At the same

time, in the background of informationization, people's information literacy and their ability to operate computers and smartphones are constantly improving, and they can collect, process, and apply information, and they have become accustomed to the "click, search, and fragmentation" method. Data mining technology has been widely used and achieved in various industries. By translating through online English translation platforms, data mining technology is used to mine and analyze translation data, and the correlation between the English translation and various factors is explored using model training results, which is of great value to translation. All of these conditions help to break away from traditional translation teaching and realize personalized translation [3]. All these conditions help to break away from the constraints of traditional translation teaching and realize personalized translation. Thus, it can be seen that an in-depth study of translation can, to a certain extent, promote the development of translation and improve the translation level. In this paper, based on a review of relevant research literature, we conduct an empirical study on the design elements and influencing factors of an English translation and explore the significance of the research on English translation teaching. With the emergence of "Internet+", all walks of life have undergone radical changes, and "Internet + translation" has become a development trend, bringing new opportunities and challenges for translation. Today's top research for data secondary mining has made great progress compared to the previous research, among which the most important algorithms are the following: concept tree boosting algorithm, association algorithm, classification algorithm, and genetic algorithm. Through the online translation English platform, we use data mining technology to analyze English translation data and establish relevant models to explore the correlation between the English translation and various elements, which is of great value to students' learning and teachers' teaching. The author has conducted an in-depth study on this. Contact between languages through translation is an important factor in triggering changes in the target language [4]. So far, the role of translation in the target language has only been seen in relatively scattered studies, lacking sufficient attention and systematic research. Accordingly, the relationship between translation and the development of English is examined more systematically by defining how translation acts on the development of English and the extent of its influence.

Data mining techniques have been widely used and achieved in various industries, and it is necessary to apply data mining techniques to translate through online English translation platforms, to mine and analyze translation data, and to explore the correlation between the English translation and various factors using model training results, which is of great value to translation [5]. Data mining is often an iterative execution process; that is, data mining needs to be performed several times. In this process, the overarching deployment planning and guidance is essential, and the data mining process generally has several stages, which are data preparation, data selection, data reduction, data mining, and knowledge evaluation. This research area

belongs to one of the aspects of data mining for English translation. According to the different application areas of data mining, translation data mining can be divided into data mining in English translation research, data mining in the English translation management process, data mining in the research management process, and so on. Data mining is to mine useful and potential knowledge from a large amount of data, and data mining is one of the core research areas of artificial intelligence and is one of the methods to solve English translation problems, which is trained by a large number of data sets, and the task of data mining is accomplished by various related algorithms. Deep translation, as one of machine translation, has made great progress in the fields of artificial intelligence and human-computer interaction, and deep translation has brought a new wave of machine translation. At present, deep translation technology has been widely used in various fields.

2. Related Works

In the development process of human history, a huge amount of information has been generated in the process of production and life, especially in the present information explosion era in which we live, the work is not very realistic if people still use human resources to deal with information [6]. However, the current database platform can realize the simple management of data, but cannot realize the secondary mining and utilization of data, and cannot aggregate the inner association existing in the information, to serve the production life. Secondly, the existing technology of data mining cannot meet the requirements; so far, there is no mature and available system tools to analyze and apply the rules to a large amount of data in a rational way, which leads to the phenomenon of "data explosion but poor knowledge" [7]. The modern human use of information no longer stays in the use of the data but needs to deep analysis of data information to get the objective laws behind the big data and to guide people to making feasible decisions in line with the law [8]. This is the intrinsic meaning of data mining. Data mining (DM), also known as knowledge discovery, is also the process of integrating simple information again using statistical analysis, machine learning, and other methods for database discovery information [9]. Today the world's top research for data secondary mining has made great progress compared to the previous research; the most important algorithms are the following: concept tree enhancement algorithm, association algorithm, classification algorithm, genetic algorithm, and so on. There has been researching in data mining since the beginning of the connection to the Internet and numerous results have been obtained in the decades of development. Nowadays, the data mining technology situation is mainly formed by the triad of data warehousing, statistical and intelligent technologies [10]. Applied data mining techniques for society and business are also still in the experimental stage and have not yet been applied on a large scale. In China, data mining technology has been studied on online English learning platforms, university English listening, teaching quality, examination results, course performance analysis, and management. In

terms of online English learning platforms, clustering analysis is used to cluster students' mastery of English levels, an association rule algorithm is used to analyze the connection existing between practice questions and practice questions, and a genetic algorithm is used to develop an automatic organization of English learning content system; in terms of university English teaching quality, data on teaching factors affecting students' learning quality are mined and analyzed according to the data in universities and modeled by decision tree algorithm [11], mining the importance of each factor affecting learning quality and improving teaching quality in a targeted way according to the analysis results; in university English listening, based on the English listening performance in a university, Internet research data, and other resources, using data such as gender, teaching means, learning purpose, test difficulty, and teaching language as input, mining the first-factor affecting listening test performance is of interest and the second factor is teaching means. The research in college English level 4 exam is based on the data of English foundation, effort, learning method, interest, English atmosphere, and so on, using algorithms such as ID3 in a decision tree to mine the rules to predict whether they can pass the exam.

Data related to English teaching and learning in colleges and universities are used. Data mining includes a variety of data classification algorithms; the common ones include decision tree classification, neural network, and genetic algorithm fuzzy set method. Among these mainstream data classification methods, decision trees are the most classical and commonly used techniques, and they are reliable data models that can be relied on in both engineering and modeling. In the computer-assisted language model, using students' feedback data on learning English online and using multivariate linear models, we mine and analyze the factors affecting students' learning attitudes; using face-to-face teaching and independent learning models, we mine students' enhanced independent learning based on students' feedback data on learning English, teacher interview student survey data, final assessment, teacher feedback, and other data and students on online English writing platform. The project will use behavioral data, such as writing time, number of submissions, content changes, number of revisions, and so on [12] to mine and analyze the factors affecting English writing and propose targeted solutions to improve English writing based on the influencing factors; using the online English learning platform, we will mine the influence of each factor on learning satisfaction through the data of questionnaires filled by students [13]. The focus of the "Digital Learning Research Project" includes student achievement, teaching and learning process, knowledge sharing, and communication. The author finds that the current application of data mining technology in university English teaching is mainly to use test results data or feedback data such as questionnaires to mine and analyze. According to students' English training process or learning path data mining and analysis, and according to the learning path to mine the factors which strongly related to test results, this paper makes a longitudinal analysis of student learning data in the university English teaching and assessment system

through the portrait and uses relevant models to mine factors related to exams.

3. Construction of a Complex System Model for English Translation Based on Web Discrete Data Mining

3.1. Network Discrete Data Mining Model Design. The process of data mining varies in different application areas and different technical approaches, and the corresponding process is developed for different mining M-targets and requirements. In general, the data mining process can be summarized into five phases: preparation phase, data preprocessing and analysis phase, model training phase, model validation and evaluation phase, and online use phase:

(1) *Preparation Phase.* In the preparation phase of data mining, the main focus is on understanding the project objectives and business, user needs, and so on and being able to define the problem as well as the initial definition of mining objectives, while data collection needs to be carried out [14]. There are various ways to collect data, such as questionnaires, user interviews, observations and questions, group discussions, and so on, each of which has more or fewer advantages and disadvantages. One of the main sources of a large amount of data in the Internet industry is logs, where information about users and various behaviors and operations are recorded accordingly, and also data can be obtained through techniques such as crawlers. The source of data in this study is the data generated by the students when they log in to the English training system for English practice and exams and this data is saved in the database.

(2) *Data Preprocessing and Analysis Phase.* In the data preprocessing and analysis phase, the collected source data are processed, and the incomplete, inconsistent, and noisy data in the source data are processed and integrated through data preprocessing techniques and methods to provide "clean" data for mining; this phase needs to clarify the mining task. In this stage, we need to clarify the mining task, whether it is a classification problem or a regression problem, and process the sample data with features according to the feature engineering method.

(3) *Model Training Phase.* The model training phase is to select a suitable machine learning model for the processed "clean" data. Mining analysis, such as classification problems, needs to choose the appropriate classification algorithm. From the perspective of translation studies, the combination of descriptive translation studies and corpus linguistic studies not only feeds the study of contact between languages through translation but also provides interesting pointers. The modeling is done by continuously adjusting the parameters to achieve the best results, and the model training often requires going back to

the previous stage and reprocessing the data and features before training.

(4) *Model Validation and Evaluation Phase.* After the machine learning model has been trained, the final results of the mining need to be analyzed and evaluated; that is, the model obtained from the training is used to test the data for effect validation. For example, the classification problem can be evaluated by accuracy, recall, and so on, and in the regression, the problem can be evaluated using the mean square error; if the validation effect does not meet the demand target, it is necessary to go back to the previous stage; sometimes need to replace the algorithm again, until the target is reached.

(5) *Online Use Stage.* In this stage, the whole data mining work has been completed and the generated model is used for online use or for solving more complex problems. The specific flow of the data mining processing model is shown in Figure 1.

Data mining is often an iterative process of execution; that is, it needs to be performed several times. In this process, overarching deployment planning and guidance is essential. The data mining process generally has several stages; these processing stages are data preparation, data selection, data reduction, data mining, and knowledge evaluation. Data mining is the result of a cycle, and its process requires the use of algorithms to organize and calculate; data algorithms are not a quick fix in the actual application of multiple adjustments to get the algorithm steps suitable for data mining. Throughout the history of machine translation development, research experts have made substantial breakthroughs in the field of machine translation and have proposed a variety of typical machine translation model frameworks. These include rule-based machine translation, instance-based machine translation, statistical machine translation, and neural machine translation. Data mining processing model with data collection is mainly based on data mining to achieve the definition of data; from other forms of data existence to extract available information, this information may be the first extraction; it may also be repeated extraction. At the first time, the discovery of data is generally selected with representative characteristics of the data, and multiple repeated extraction is often the basis of the last data selection to do data screening. This greatly saves time in data organization. Knowledge discovery algorithm points to determining or designing the process of the algorithm [15]. The application of the algorithm is purpose-specific and therefore needs to be based on the object and rediscovery to get the data set. Data mining: data filtering is carried out by adopting established algorithms for different needs in the context of knowledge discovery. Pattern feedback is then used to determine the feasibility of the mining process based on the identification of the results of the target data. If the data meet the requirements, then regress to the data collection stage and use such algorithms on a large scale; if the target data results do not meet the mining requirements, the algorithm

is reselected; if the data results are close to the target requirements, the algorithm parameters are changed directly back to the algorithm stage. For the data screening process, it is necessary to select the utilized data for the operation to avoid the waste of time, while the data at each stage is checked and the algorithm is modified in real time, and the screening and application of the target data are completed under the continuous modification of the algorithm. The data mining model of the architecture is shown in Figure 2.

The process of data mining is required to target a certain domain, so the whole process requires relevant expertise. Nowadays, the data application areas are financial markets, biotechnology, and Internet technology. When you shop in Taobao, the traces of your browsing are analyzed by big data and then similar product pages are pushed on your homepage. The above analysis can prove that data mining technology is gradually applied to the various public life of people and it shows more obvious advantages. We can intuitively see that this method is only suitable for the case of relatively uniform distribution of data points, without considering the sparsity of the discrete points themselves, and when the data points are not uniformly distributed, the cumulative chord length parameterization method can be adopted, and the corresponding parameterization formula is

$$K = \{i \times d\},$$
$$D = \frac{M_1}{n - p}. \tag{1}$$

Data mining itself is a comprehensive interdisciplinary application derived from information technology in various industries, so data mining is used in various industries with different algorithms. Different approaches are also required in targeting a specific problem, and different techniques and methods are used to get suitable algorithms and steps. Data Mining Frontiers has made a comprehensive compilation of the selection of techniques and tools for the application of data [16]. Only data mining methods related to the dissertation research are detailed below. Data mining includes a variety of data classification algorithms; the common ones include decision tree classification, neural networks, and genetic algorithm fuzzy set methods. Among these mainstream data classification methods, decision trees are the most classical and commonly used techniques; both in engineering and modeling, decision trees are reliable data models that can be relied upon. The latter comprehensive rating system mainly uses the strategy tree algorithm. The following is a brief description of the decision tree classification method decision tree is a tree structure with a root node on which multiple subnodes are distributed. Here each internal node represents a single attribute of the test, while each branch corresponds to the test output and each leaf node represents the class as well as the distribution of the class. The topmost of these nodes is the root node. It represents the process of class selection by the students and it predicts the characteristics of class selection by the students. The internal nodes are represented by rectangles while the leaf nodes are represented by diamond shapes. In practical
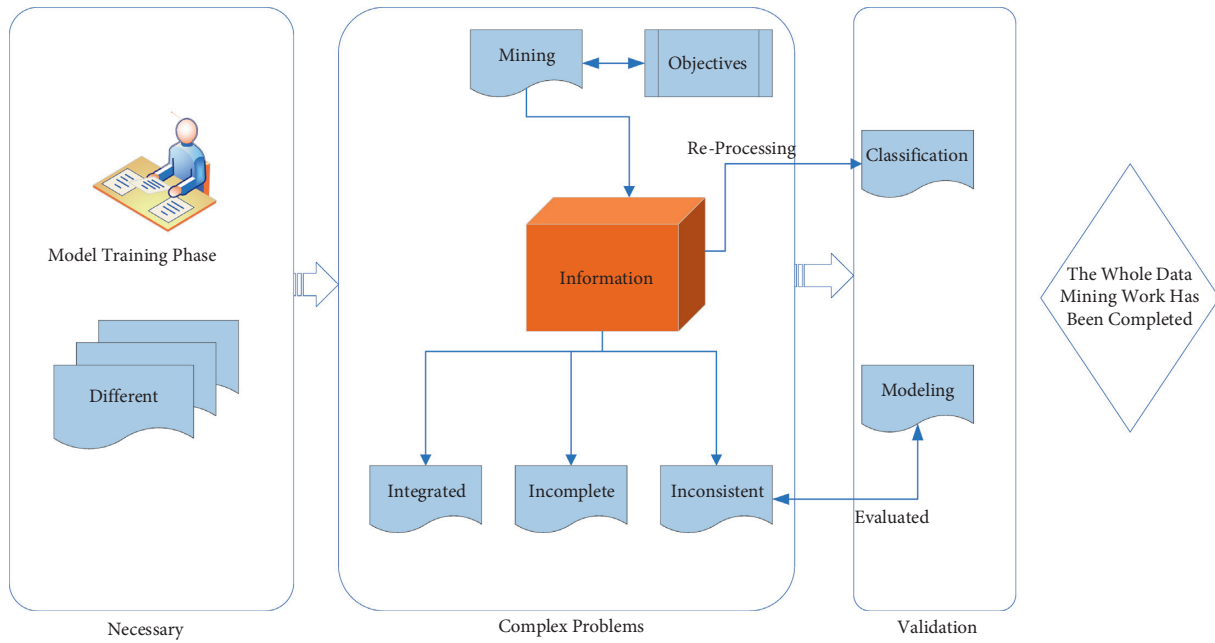
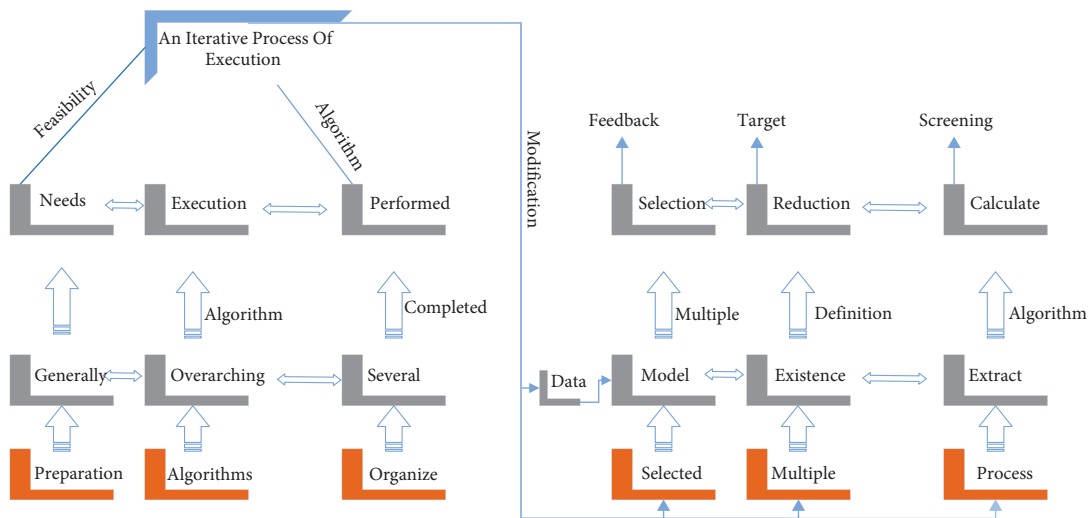FIGURE 1: Specific flow of the data mining processing model.

FIGURE 2: Data mining model architecture.

application, according to the specific attribute values on the decision tree, each node should be connected with the algorithm to form a unified algorithm. Each node on the decision time, except the end node, is a specific problem of a decision, traversing to a node will choose "no" or "yes", to get different results, and then judge the data selection of the final solution. Each path of the decision tree is a desirable rule, so it can be directly converted into a classification rule. This difference is very obvious when the dataset is not loaded into the main memory. The time required for record access is given, and when a partial classification model has been built, the data scan is just a slice of the entire training data.

$$L = \sum \log P(y, x). \qquad (2)$$

3.2. Complex System Model Construction for English Translation.

One of the reasons why the impact of translation on the target language has long been neglected is that the extent of its impact is difficult to measure using valid tools and instruments. The development of the discipline of historical languages, which is closely linked to historical sources and empirical evidence, cannot survive and thrive on intuitive and introspective data alone [17]. In terms of the development of historical linguistics, each innovation in research methods and techniques has opened up finer areas for its development. From the perspective of translation studies, the combination of descriptive translation studies and corpus linguistic studies not only feeds the study of contact between languages through translation but also provides interesting pointers. For example, translated texts

have both their commonalities and their unique linguistic typological, textual, and pragmatic characteristics due to their origins in another language, and the changes in the relationship between these translated texts and the original text, and their intrinsic mechanisms of operation, can be better interpreted in the perspective of historical linguistics. This study mainly attempts to analyze the linguistic features of translated texts in three time periods, namely, the 1930s, 1960s, and 1990s, through the construction of a new type of ephemeral composite corpus, that is, an ephemeral corpus combining parallelism, analogy, and reference, which includes literary corpora and to examine the continuity between the English source text, the Chinese translated text, and the Chinese original text. Accordingly, how translation acts on the development of modern Chinese and the extent of its influence are defined to further examine the relationship between translation and the development of modern Chinese and to move the corpus chronological research in this field forward. A diagram of the framework structure is shown in Figure 3.

Machine translation, which uses computers to automatically translate the source language into the target language, bridges communication between people from different civilizations. Throughout the history of machine translation development, domestic and foreign research experts have made substantial breakthroughs in the field of machine translation and have proposed a variety of typical machine translation model frameworks. These include rule-based machine translation, instance-based machine translation, statistical machine translation, and neural machine translation. The rapid growth of data volume has put forward new requirements for data query, data storage, and data management, and there is an urgent need for a new tool to intelligently mine valuable information and knowledge from the massive data. This section will focus on the statistical machine translation model and neural machine translation model frameworks that are mainstream today and used in subsequent chapters in detail. Phrase-based statistical machine translation models use phrases as the smallest translation unit. The basic idea is that the source and target language sentences are divided into phrase sequences and the source language phrases are translated into the target language using the phrases as the basic translation units. More specifically, the statistical machine translation model takes the bilingual aligned corpus as input, generates phrase translation rules from the source language to the target language through the bilingual aligned corpus, and finally produces the final target language translation through the sequencing model and language model. In general, statistical machine translation models consist of three main components: translation models, language models, and decoding algorithms.

$$P\left(S_1\right) = P\left(S_{i-1}, S_{i-2}\right). \tag{3}$$

There are two types of approaches to the study of translated texts: the comparative linguistic approach and the descriptive translation research approach. The former emphasizes the differences between the language systems of the source and target languages and how these differences affect the translation process and the translation outcome; the latter focuses mainly on the study of the translation process itself. Descriptive translation research focuses on the translation itself, and it addresses the question of "why the translator translates in this way" rather than "how to translate." The text-oriented empirical methodology of corpus linguistics fits well with the study of descriptive translation, and therefore the study of commonalities in translation has been associated with descriptive translation research from the beginning. Contrastive linguistics is concerned with both the inherent characteristics of a language and its comparability, although it is more concerned with interlinguistic differences. These two research methods complement each other [18]. Contrastive linguistics and descriptive trans linguistics can draw on each other's approaches, in that they both examine how "the same thing" is expressed in another way, albeit for different purposes. In the past two decades, corpus translation studies have been dominated by descriptive translation studies, and the linguistic contrast model has been neglected. While elevating the status of the translated text as an integral part of the cultural system of the translated language, descriptive translation studies have overlooked the fact that translated texts are derivative bodies with dual or multiple complexities that are characteristic of the source language itself. The translation is the product of two texts and two sociocultural events and is "intertextuality" in nature. The overuse of any one research method may lead to its inextricable trap. In the background of information technology, people's information literacy, computer, and smartphone operation ability are increasing, and they can collect, process, and apply information, and they are accustomed to "clicking, searching, and fragmentation"; all these conditions help to get rid of the traditional translation teaching and realize personalized translation. The comparative study of the manifestation phenomenon in this paper is mainly an English language study, which focuses on the comparison of linguistic systems rather than on the eagerness to make generalizations and deductions about linguistic commonalities. This is one of the mistakes that corpus translation studies make, and it is this mistake that the study of conspicuousness suffers from, and it is superficial and harmful to categorically attribute it to translation commonalities before careful analysis. Accordingly, this paper combines the research method of comparative linguistics with the study of descriptive translation from the perspective of the English and Chinese languages, focusing on the differences in language systems to examine whether manifestation and excessiveness present the same picture in different historical periods or whether they change in stages, instead of conducting a top-down study from translation commonalities only. The study of the relationship between translation and language development has mostly been conducted in the previous literature along two lines: the perspective of translation studies and the level of historical linguistics. The former tends to overemphasize translation commonalities at the expense of its possible effects; the latter has never attempted to compare the typical features of translated texts
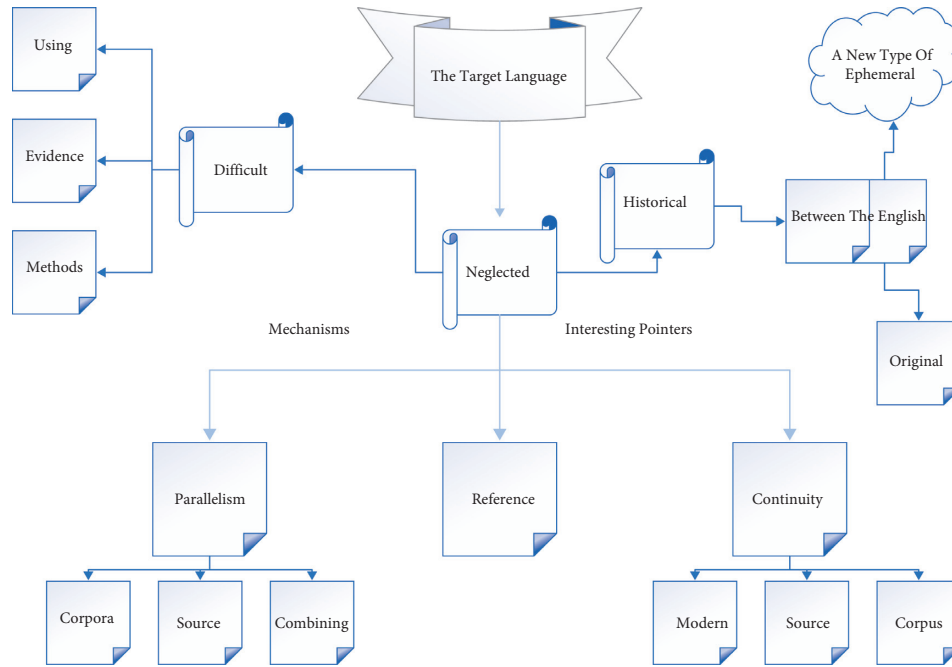
FIGURE 3: Framework structure diagram.

with nontranslated texts and thus explore language change. This paper attempts to combine these two research approaches to further explore what effects the features of translated texts may have on the developmental changes of the original text from the perspective of language contact. If features such as manifestation or simplification do exist in the translated text, then these features are bound to have an impact on the original text as well. The schematic diagram is shown in Figure 4.

The examination part of the parallel corpus focuses on the translation sources of these two types of manifestation phenomena, dividing them into two categories: source language equivalent components and source language nonreciprocal components. They are mainly annotated and counted by randomly sampling 500 English-Chinese language pairs for each period. According to the different application fields of data mining, translation data mining can be divided into the application of data mining in English translation research, the application of data mining in the English translation management process, and the application of data mining in the research management process. Among them, source language equivalent components are divided into direct correspondence and adverbial or prepositional correspondence; source language nonreciprocal components mainly include four categories of interlingual additive explicit, alternative explicit, omitted implicitly, and alternative implicit. Connective manifestation is divided into three categories according to different semantic relations: reciprocal, contrastive, and causative relations and a case study is conducted for each category separately; denotative manifestation is divided into three parts: object pronouns, personal pronouns, and manner-indicative pronouns. Among them, the neutral pronoun is mainly used as a case study. In the parallel examination, we take the manifestation
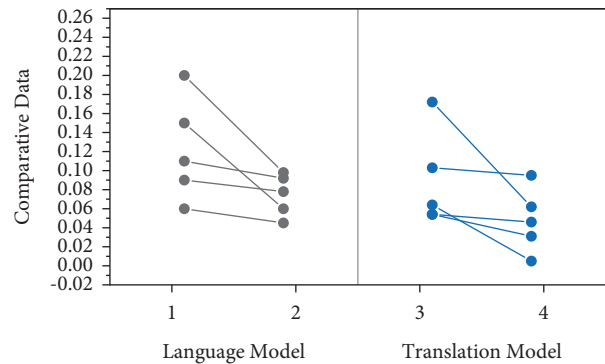

FIGURE 4: Architecture diagram.

indicator in the translated language as the search term for its corresponding component in the English source language and compare its ephemeral changes in three periods to determine whether it is influenced by the English source language's transitivity and what is its degree of change; the analysis of the manifestation and implicitization of its nonreciprocal component can further examine the translation commonality and whether it will be in different ephemeral stages.

4. Analysis of Results

4.1. Analysis of Discrete Data Mining Models for Networks. For decision tables and rule inference, ART is in some ways much faster than other methods due to the different search methods. Whereas previous algorithms obtained only one rule per scan, ART obtains one rule set at a time, and to reduce the number of scans of the dataset, ART must estimate candidate solutions. This difference is very obvious

when the dataset is not loaded into the main memory [19]. Given the time required for single-record access, when a partial classification model has been built, the data scan is just a slice of the entire training data. The plain Bayesian classifier can be used as a reference because its I/O cost is optimal, and it only has to perform one scan of the entire data set. To complete the experiment, it is also necessary to calculate the number of disk pages read by each algorithm for pages of different sizes. Since RT, CN2, and RIPPER are iterative algorithms, the number of disk pages read by these algorithms decreases proportionally with the disk size. This is because it is impractical to compute the arc length of the original curve from a discrete set of points with noise. However, although the cumulative chord length parameterization is more accurate than the uniform parameterization, it still has large deviations from the data point parameter values, especially when the data is noisy or when the data point distribution is too sparse. In the past, such errors could not be reduced because the known set of data points only contained the location coordinates of the data points. In contrast, this paper can obtain a more accurate parameterization with the help of a known geometric feature, the normal vector. In the following, we present an improved method for parameterizing discrete points based on normal vector information. Data mining has been studied since the beginning of the connection to the Internet, and numerous results have been obtained in the decades of its development. Nowadays, the data mining technology situation is mainly formed by the triad of data warehousing, statistical and intelligent technologies.

$$t_i = \frac{[p_i - p_{i+1}]}{d}. \tag{4}$$

According to the sample data in the input classifier, the variables that are strongly associated with predicting students' passing the exam, that is, the main factors that affect students' passing the exam are scores in each question type, students' exam scores, homework completion, number of English training sessions, interest in English, and fourth- and sixth-grade status, are summarized and analyzed for each factor as follows: students' scores in each question type have the greatest impact on students' passing the exam; when a student has no particular weakness in listening, speaking, reading, and writing, the probability of the student passing the exam is very high; when the student can reach the passing level in all questions, the probability of the student passing the exam is also high; in addition, students who passed the fourth and sixth levelshave a higher probability of passing the exam. From the data, we can see that both the scores of each question type, the final scores, and the grades of levels 4 and 6 can, to a certain extent, reflect the students' original English foundation, so the students' original English foundation has a great influence on passing the exam. The number of times students completed homework and participated in English training had a relatively large impact on passing the exam, and the probability of students passing the exam was higher if they completed homework at a high rate or participated in

English training more times in the experiment. Therefore, a high rate of completion of homework and a high number of training sessions has a relatively high impact on passing the exam. Students' interest in English has an impact on the prediction of passing the test, and the experimental prediction shows that the more interested students are in English, the more likely they are to pass the test. In the modern "student-centered, teacher-oriented" teaching theory, the importance of cultivating students' interest in learning is also emphasized, through cultivating students' interest in learning, not only can stimulate students' learning motivation and mobilize students' learning enthusiasm, but also can promote students' active participation in learning and cultivate students' sense of self-directed learning. Interest in English can, to a certain extent, reflect students' emotional and psychological factors in the process of learning English. For example, in speaking, some students' English pronunciation is affected by geographical factors, such as local accent, which leads to no one being able to understand them when they express themselves orally so that in the long run, students will no longer be willing to speak English actively or even hate learning English. Therefore, the level of interest in learning English can have a greater impact on the effectiveness of English learning. From the data results, we can see that the gender factor also has a greater impact on English performance. By observing the sample data and analyzing it, we found that in the case of the equal number of girls and boys, the number of girls passing the exam is more than boys, and the probability of girls passing the exam is greater than that of boys in the predicted results, which is found to be related to the fact that female students are more positive and active in their learning attitude than male students. The variables "difficulty of questions" and "weight of questions" in the sample do not have a large range of variation, the difficulty is moderate, and the weight of each type of question does not fluctuate much, so the prediction result is not very influential. Data mining processing model of specific data collection is mainly based on data mining to achieve the definition of data, from other forms of data existence to extract available information; this information may be the first extraction; it may also be repeated extraction. The first discovery of data is generally selected with representative characteristics of the data, and multiple repeated extraction is often the basis of the last data selection to do data screening. This greatly saves the time of data collation. Higher education institutions are under great competitive pressure, and they need to promote the development of various aspects and create a better environment for students' learning and life in a reasonable way, and to achieve this goal, they can use the way of joint schooling. For example, our school has cooperated with enterprises such as the National Investment Potash Company, a clothing design company, and a machinery repair factory. In addition, the evaluation process needs to take into account fairness and rationality and exclude the influence of human factors; at this time, it is necessary to use the classification model algorithm of data mining, that is, the use of decision tree model for English translation

processing, to make a more objective evaluation of its performance. The experimental analysis is shown in Figure 5.

In the method proposed in this paper, a lexical-level systematic fusion technique is used, and the main steps are as follows: first, one of the target language translations from multiple models is selected as the lexical-level alignment reference, and the translations from other models are aligned to this alignment reference, and a confusion network is built through the word alignment information of each model translation; then, the words in the confusion network are estimated in terms of their disposition; finally, the disposition combined with the language model, length penalty, and other features in the statistical machine translation model is decoded, and finally the translation with the highest decoding score is selected as the reference translation for the final model fusion. Our proposed method based on pairwise learning multimodal fusion avoids the degradation of performance due to error propagation from multiple translation transformations and steadily improves translation performance to a certain extent. To investigate the impact of the two modules in our proposed method on the final translation performance, we conducted multiple sets of experiments separately.

$$R = \log P(y) = \sum_i \log P_y(y - y_1). \tag{5}$$

4.2. English Translation System Modeling Application Implementation. It mainly discusses the research on the application of data mining related algorithms in modeling English translation systems, from three perspectives of association rules, clustering rules, and classification rules, the algorithm suitable for this case is selected to mine the hidden knowledge rules in English translation, and the code implementation of Apriorism algorithm and K-means algorithm in python language is chosen for the translation data in the case. Useful conclusions were obtained after running. These conclusions will provide a favorable reference and scientific basis for curriculum construction, professional construction, and improvement of teaching quality. For accuracy comparison, see Figure 6.

A new generative adversarial network model is proposed: multiscale discriminative generative adversarial network. Traditional generative adversarial network algorithms generate translations with many unreasonable falsified objects, and the multiscale discriminative generative adversarial network algorithm can well avoid this problem by discriminating multiple scales of English translations simultaneously. One of the reasons why the impact of translation on the target language has long been neglected is that the extent of its impact is difficult to measure using valid tools and instruments. The development of the discipline of historical languages, which is closely linked to historical data and empirical evidence, cannot survive and thrive on intuitive and introspective data alone. Based on this, a weight-sharing multiscale discriminator is proposed. In addition, a translation algorithm based on generative adversarial networks is proposed, which can improve translation efficiency.
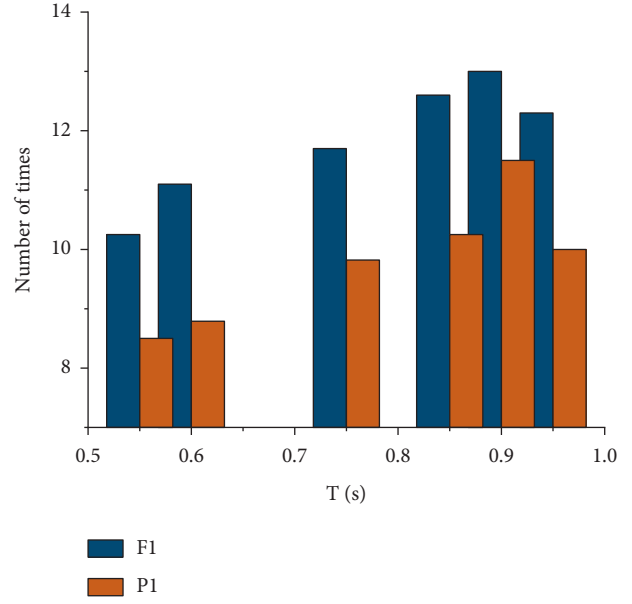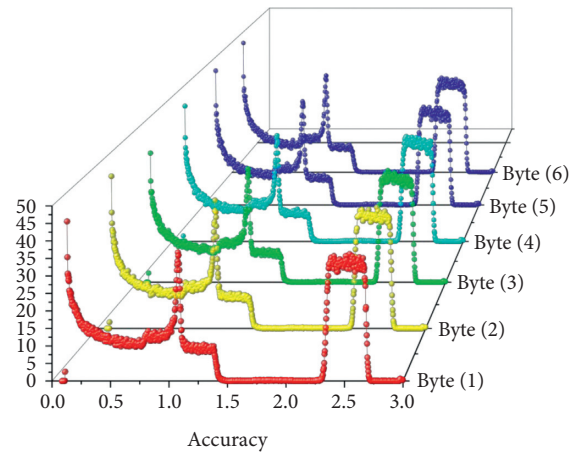

Figure 5: Experimental analysis diagram.

Figure 6: Accuracy comparison.

Finally, the multiscale discriminative generative adversarial network (MSD-GAN) is applied to English translation, and the experimental results verify the effectiveness of the algorithm.

$$\mathrm{Inf}(S_1, S_2) = \sum P^2 \log_2 P. \tag{6}$$

We explore the unsupervised neural machine translation model and propose a new framework for unsupervised neural machine translation for both long-distance and near-distance language pairs. Experimental results show that our proposed method can perform unsupervised machine translation when there is no parallel corpus, but for long-distance language pairs, better cross-linguistic word embedding methods need to be found to improve the overall model performance [20]. A neural machine translation model based on pivotal languages and a new framework for neural machine translation of pivotal languages based on pairwise fusion are proposed. The experimental results show
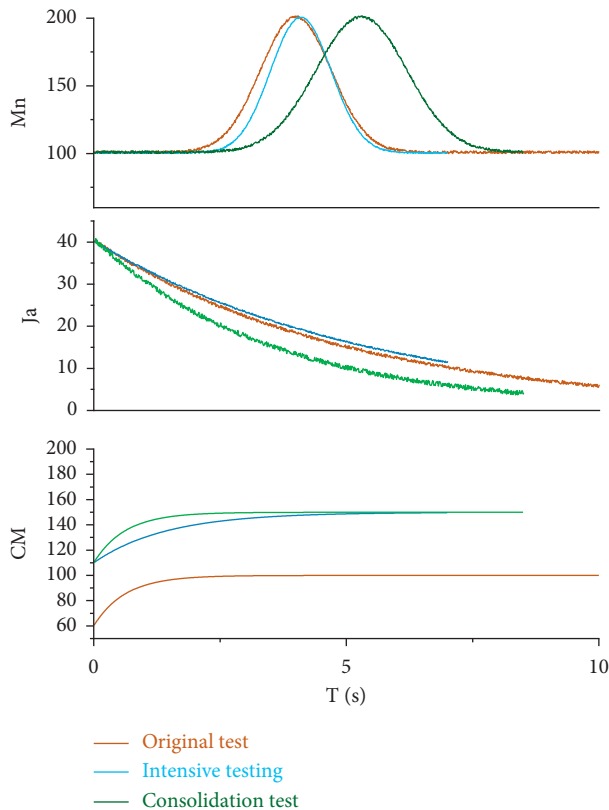
FIGURE 7: System modeling application implementation.

that our proposed approach achieves enhancements compared to the traditional English translation system modeling. The system modeling application implementation is shown in Figure 7.

5. Conclusion

With the continuous improvement of computer performance, data mining has also been vigorously developed in recent years, and data mining-based research systems, especially in the open field, have achieved impressive results. Data mining technology has gained wide attention from the information session since it was proposed and researched, and it is mainly application-oriented and has achieved great success in these applications, and as an emerging data analysis technology, it has been successfully applied to business fields such as finance, sales, transportation, insurance, and telecommunications. With the booming development and demand, English translation is bound to grow, and these will impose higher standards and requirements on English translation. Based on the relevant attributes generated by the system modeling and the data of English translation, the theory, and technology related to data mining are studied in-depth, and the construction of this system is based on the application in the modeling of the English translation system, and the discrete data technology is utilized for the design, which has a certain cutting-edge overall, which is helpful to improve the management level of English translation and as well as the professional quality of

English translation. For the current development of English translation, this system can be better applied in various fields of translation.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] M. S. Mahdavinejad, M. Rezvan, and M. Barekatain, "Machine learning for internet of things data analysis: a survey," *Digital Communications and Networks*, vol. 4, no. 3, pp. 161–175, 2018.

[2] H. Jelodar, Y. Wang, C. Yuan et al., "Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey," *Multimedia Tools and Applications*, vol. 78, no. 11, pp. 15169–15211, 2019.

[3] F. A. Abad and A. A. C. L. Chaparro, "Data-mining techniques in detecting factors linked to academic achievement," *School Effectiveness and School Improvement*, vol. 28, no. 1, pp. 39–55, 2017.

[4] Q. Hou, M. Han, and Z. Cai, "Survey on data analysis in social media: a practical application aspect," *Big Data Mining and Analytics*, vol. 3, no. 4, pp. 259–279, 2020.

[5] H. Jelodar, Y. Wang, R. Orji, and S. Huang, "Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 10, pp. 2733–2742, 2020.

[6] D. H. Maulud, S. R. M. Zeebaree, and K. S. S. Jacksi, "State of art for semantic analysis of natural language processing," *Qubahan Academic Journal*, vol. 1, no. 2, pp. 21–28, 2021.

[7] J. Lara-Rubio, A. F. Villarejo-Ramos, and F. Liébana-Cabanillas, "Explanatory and predictive model of the adoption of P2P payment systems," *Behaviour & Information Technology*, vol. 40, no. 6, pp. 528–541, 2021.

[8] Y. Liu, D. Liu, Y. Zhang et al., "Markov clustering analysis-based validation for traditional Chinese medicine syndrome patterns of poststroke cognitive impairment," *Journal of Alternative & Complementary Medicine*, vol. 25, no. 11, pp. 1140–1148, 2019.

[9] R. Lambiotte, M. Rosvall, and I. Scholtes, "From networks to optimal higher-order models of complex systems," *Nature Physics*, vol. 15, no. 4, pp. 313–320, 2019.

[10] M. W. Rodrigues, S. Isotani, and L. E. Zárate, "Educational Data Mining: a review of evaluation process in the e-learning," *Telematics and Informatics*, vol. 35, no. 6, pp. 1701–1717, 2018.

[11] T. Miranda, L. R. Sousa, A. T. Gomes, J. Tinoco, and C. Ferreira, "Geomechanical characterization of volcanic rocks using empirical systems and data mining techniques," *Journal of rock mechanics and geotechnical engineering*, vol. 10, no. 1, pp. 138–150, 2018.

[12] M. Sudha and A. Kumaravel, "Comparative Analysis between Rough set theory and Data mining algorithms on their prediction," *Global Journal of Pure and Applied Mathematics*, vol. 13, no. 7, pp. 3249–3260, 2017.

[13] V. Salehi, B. Veitch, and D. Smith, "Modeling complex socio-technical systems using the FRAM: a literature review,"

Human Factors and Ergonomics in Manufacturing & Service Industries, vol. 31, no. 1, pp. 118–142, 2021.

[14] Y. Mohammadnian, T. Amraee, and A. Soroudi, "Fault detection in distribution networks in presence of distributed generations using a data mining-driven wavelet transform," *IET Smart Grid*, vol. 2, no. 2, pp. 163–171, 2019.

[15] K. Zhang, C. Shen, H. Wang, Z. Li, Q. Gao, and X. Chen, "Cluster computing data mining based on massive intrusion interference constraints in hybrid networks," *Cluster Computing*, vol. 22, no. 3, pp. 7481–7489, 2019.

[16] S. Wu, J. Liu, and L. Liu, "Modeling method of internet public information data mining based on probabilistic topic model," *The Journal of Supercomputing*, vol. 75, no. 9, pp. 5882–5897, 2019.

[17] J. Gu, F. Fu, and Q. Zhou, "Penalized estimation of directed acyclic graphs from discrete data," *Statistics and Computing*, vol. 29, no. 1, pp. 161–176, 2019.

[18] D. M. Khan, A. U. Rehman, M. U. Rehman, F. Shahzad, and N. Saher, "A discrete formulation of a unified data mining model," *The Nucleus*, vol. 57, no. 4, pp. 135–140, 2021.

[19] V. A. Sergeev and N. A. Korgin, "Identification of integrated rating mechanisms as an approach to discrete data analysis," *IFAC-Papers Online*, vol. 54, no. 13, pp. 134–139, 2021.

[20] M. Jeihouni, A. Toomanian, and A. Mansourian, "Decision tree-based data mining and rule induction for identifying high quality groundwater zones to water supply management: a novel hybrid use of data mining and GIS," *Water Resources Management*, vol. 34, no. 1, pp. 139–154, 2020.