Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 6803802, 12 pages
https://doi.org/10.1155/2022/6803802

Research Article

Construction of Computer English Corpus Assisted by Internet of Things Information Perception and Interaction Technology

Miao Tian

School of Foreign Languages, Weinan Normal University, Weinan 714099, Shaanxi, China

Correspondence should be addressed to Miao Tian; tianmiao@wnu.edu.cn

Received 9 August 2022; Revised 20 September 2022; Accepted 27 September 2022; Published 11 October 2022

Academic Editor: Jun Ye

Copyright © 2022 Miao Tian. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Aiming at the situation that multiple sensing devices in the sensing layer sense the same sensing quantity, multiple redundant measurement data will be generated. In this paper, an improved Kalman filtering algorithm is used. The algorithm combines with numerical test theory to remove abnormal error data before filtering, thereby reducing the error of filtering result and improving its reliability. The parallel multiprotocol processing mechanism of the gateway is designed, and the IoT gateway has a hierarchical modular processing design for different protocol signals of the heterogeneous perception layer. The web interaction function with good platform versatility is designed to provide a visual interaction interface for the application layer, which is convenient for the application layer terminal to monitor and manage the perception layer equipment. This paper realizes the retrieval and query function of keywords, wrong sentences, and English character errors. This corpus enriches the less existing Madagascar corpus and establishes a separate English character error query database. This paper describes the overall process of corpus from preliminary preparation to establishment and the related use instructions of corpus, makes a statistical analysis of the current corpus data, and puts forward some suggestions for building a computer English corpus according to the analysis results.

1. Introduction

Situational awareness is one of the main features of IoT intelligence. The introduction of situational awareness technology enables edge smart devices to perceive situational changes in dynamic environments and to make intelligent decision-making services based on the perceived situation [1]. However, the complexity, dynamics, and distribution of the IoT edge environment have brought great challenges to the application of context awareness technology in the edge environment, such as high-level context awareness, context awareness under incomplete context information, and other research issues [2]. High-level situational awareness supports the identification of complex situations and realizes complex situational awareness mainly by obtaining high-level situational information from low-level situational information. In high-level situational awareness, obtaining complete low-level situational information is the basis for realizing complex situational awareness [3]. However, due to the characteristics of the IoT edge environment and the noise, conflict, unreliable links, and accidental damage in real life, data loss in wireless sensor networks is common, resulting in incomplete low-level contextual information usually obtained. The rule-based context awareness method is the main method to realize complex context awareness in the edge environment [4].

Intelligent interaction technology is an important part in the field of computer science and control scientific research [5]. The current research on intelligent interaction technology mainly focuses on the human-centered, PC-based, and traditional mobile computing "human-computer interaction," while the research on "object-object interaction" is mainly based on traditional wireless sensor networks. With the advent of the Internet of Things era, the three-dimensional world of human society, information space, and physical world will be fully connected and integrated. The interaction is developing towards an all-round and three-dimensional natural interaction, and the interaction mode also needs to be expanded from a single human-computer interaction mode to a three-dimensional world interaction mode of human, machine, and object [6]. Therefore, studying the key technologies of intelligent interaction in the Internet of Things environment and solving the interaction problems in the Internet of Things environment have important theoretical significance and urgent reality for realizing the comprehensive connection and integration of the three-dimensional world of human society, information space, and physical world [7].

Through the analysis of IoT context-aware services, this paper constructs the IoT edge context-aware service framework, which supports the reasoning of complex scenarios, and introduces a publish/subscribe-based lightweight data communication mechanism to implement services in the edge environment. After analyzing the characteristics of context fusion in the intelligent interactive system under the Internet of Things environment, on the basis of the detailed analysis of the current information fusion level, information fusion structure, and information fusion method, rule management based on rough set theory is introduced. And the generation algorithm and matching algorithm of the rules are described in detail.

The perception layer network is composed of many heterogeneous sensor networks and other device networks. It is characterized by a large number of perception terminals and a very large data traffic. The transmission of huge data traffic will block the network and increase the delay. At the same time, the transmission of these data also requires consuming a lot of energy. In this paper, two methods of data classification and aggregation and data fusion processing are proposed to improve the transmission efficiency of network data in the perception layer and reduce unnecessary data transmission such as redundancy. Among them, data classification and aggregation is a method of reducing the transmission of the same packet header data in the network by aggregating a large number of data packets into a large data packet that aggregates a large amount of useful data information, which improves the transmission ratio of useful information in the network. Data fusion is aimed at the redundant data transmission of multiple sensing layer devices to the same detection object. It performs data fusion processing before the data are transmitted to the IoT gateway. After removing redundant information, the fusion result is uploaded to the Internet of Things. This paper proposes a data fusion method that combines numerical analysis and Kalman filter fusion. It can be seen from simulation verification that this method improves the accuracy and reliability of data fusion compared with the ordinary Kalman filter fusion method.

2. Related Work

In the learning-based situational perception method, relevant scholars analyze acceleration sensor data through the backpropagation neural network to perceive human behavioral activities, such as walking, running, sitting, and other basic behaviors [8]. Relevant scholars use deep learning neural networks to analyze mobile sensor data to identify the situational patterns of human activities [9]. In the rule-based situational awareness method, the researchers proposed a situational awareness system based on ontology and rules. The system uses the ontology representation method to model situational information, recognizes complex situational patterns through rule reasoning, and is used for home care scenarios. Relevant scholars have realized situational awareness based on the combination of ontology knowledge and rules, acquired situational knowledge through rule reasoning, and applied them to the centralized environment, distributed environment, and mobile environment, respectively [10].

Related scholars have proposed a rule-based matching strategy to perceive the situation in the smart home environment and provide personalized customized services according to the current situation [11]. In the hybrid context perception method, relevant scholars use neural networks and ontology to build a framework for intelligent IoT systems and obtain low-level contextual information from sensor data through learning-based contextual awareness methods [12]. To achieve complex situation awareness through ontology reasoning, the proposed framework is mainly deployed on cloud servers to provide situation awareness services.

Investigating the research status of IoT situational perception methods shows that learning-based situational perception methods usually use raw perception data to identify human behavior and other situational patterns. Rule-based situational awareness methods are usually combined with ontology modeling methods to identify complex situational patterns through rule-based reasoning. The hybrid situational awareness method mainly integrates the above two methods [13]. The learning-based situational perception method is used to obtain low-level situational information, and based on the low-level situational information, the perception of complex situations is realized through rule reasoning. Existing research studies on perception methods are mainly deployed on cloud servers to provide context-aware services [14].

At present, due to the development of sensing technology, context can be obtained more easily, and the types and quantity of context are also more abundant [15]. The research and utilization of context-aware computing has become very important, and current research mainly focuses on context acquisition, context modeling, context information processing, and building context-aware frameworks [16].

Context modeling means that because of the existence of various contexts in the environment, by establishing a unified abstract logical model for the acquired various context information, the context information is easy to express, reason, and share.

Currently, there are a variety of context modeling methods, each with its own advantages and disadvantages, and different emphases. Context information processing is a very critical part of context-aware computing. The collected low-level context information may be incomplete and uncertain and needs to be processed to generate high-level context information to meet the needs of upper-level users.

Relevant scholars concluded that the comprehensiveness and systematicness of corpus are particularly important in the construction of English interlanguage corpus [17]. They emphasize the collection of corpora of learners at different levels of English proficiency and from different countries, not only to comprehensively collect written corpora of general students but also to track a certain number of specific students. This allows for a horizontal or vertical study of student acquisition, as well as for large-scale studies of general students or case studies focusing on specific students. Then, the scale and quantity of corpus should be ensured; the corpus samples should be evenly distributed and representative. Relevant scholars also believe that the content of corpus labeling must be comprehensive, and relevant language phenomena should be labeled at various levels such as characters, words, phrases, sentences, articles, style, semantics, pragmatics, and punctuation so as to ensure the corpus [18].

Relevant scholars suggest that it is more appropriate to use the labeling method of "correct information + biased information" for interlanguage corpora [19]. It is believed that it is necessary to focus on the processing of semantic and pragmatic annotation and to further process the text and style. In the aspect of labeling method, the concept of "limited one error and multiple labels" is proposed, and the combination of "human labeling machine-assisted" and "machine labeling human-aided" is used to try computer automatic labeling. And it is emphasized that an independent corpus for labeling and processing English character errors is established. In addition, in terms of word sense tagging, related scholars introduced several influential word sense tagging corpora that have been built from the aspects of corpus selection, dictionary selection, labeling scale, and labeling quality and proposed the bootstrapping method [20, 21].

3. Methods

3.1. IoT Edge Context-Aware Service. According to the analysis of the implementation of context-aware service at the edge of IoT in the Introduction, the edge of IoT mainly adopts the context-aware method based on rules and ontology to realize the context-aware service. There are certain limitations in information, and there is a lack of hierarchical division of situational awareness. Therefore, combining the ability of learning-based context awareness methods to process raw perception data of big data types and the advantages of rules and ontology-based context awareness methods with rich semantics, the IoT edge context-aware service framework designed in this paper adopts a hybrid context perception method to implement context-aware services. In addition, for the resource-constrained IoT edge environment, the framework also adopts a publish/subscribe-based lightweight data communication mechanism to realize the communication interaction between edge services.

The IoT edge context-aware service framework designed in this paper is shown in Figure 1. The framework mainly includes four modules: low-level situational awareness module, high-level situational awareness module, situational event processing module, and data communication module. In this framework, the realization of the context awareness method is mainly completed through the low-level context awareness module and the high-level context awareness module.

The hybrid situational awareness method can identify situational patterns such as human behavior and activities and can achieve high-level situational awareness through rule-based reasoning, that is, reasoning that supports complex situations.

According to the hierarchical division based on the complexity of the perceived situation, the realization of situational awareness is divided into low-level situational awareness and high-level situational awareness.

Low-level situational perception method: the low-level situational perception is mainly realized through the learning-based situational perception method, the basic behavior of the human body is identified by means of the deep learning method, and the behavior analysis mechanism is completed by the members of the research group. This section mainly introduces the implementation of high-level situational awareness.

The construction of contextual knowledge base is the basis for realizing high-level contextual awareness. Ontology-based contextual modeling has the characteristics of knowledge sharing and reuse, and ontology knowledge has rich semantics and supports complex contextual reasoning.

Therefore, the ontology representation method is used in the framework to model the IoT edge application field, and combined with the basic situation information, high-level situational awareness is realized through rule reasoning.

The perceptual data flow of the context-aware approach is shown in Figure 2. After receiving the sensing data of the deployed sensors, the data are divided into environmental data and behavioral sensing data through the data preprocessing stage, the behavioral features are extracted in the low-level situational perception, and some environmental data are fuzzed to obtain a low-level situation. *Situational Information*. Then, the relevant situational knowledge is extracted according to the situational knowledge base, and the rules are used to conduct situational reasoning on the basis of the basic situational information model.

3.2. Data Communication Mechanism. Adopting the publish/subscribe model to realize the cooperative interaction between services can well adapt to the dynamic changes of the environment, but it is not suitable for the resource-constrained IoT edge environment. The services in the edge environment need a lightweight data communication mechanism to realize the interaction between multiple edge services, so as to meet the complex situational events that require the coordinated processing of multiple services.

The message queuing telemetry transport protocol (MQTT protocol) is a lightweight communication protocol that supports real-time asynchronous communication and incorporates a simplified publish/subscribe model. Its main features include the following three points:

(1) One-to-many sending of messages is realized through the publish/subscribe mode, which reduces

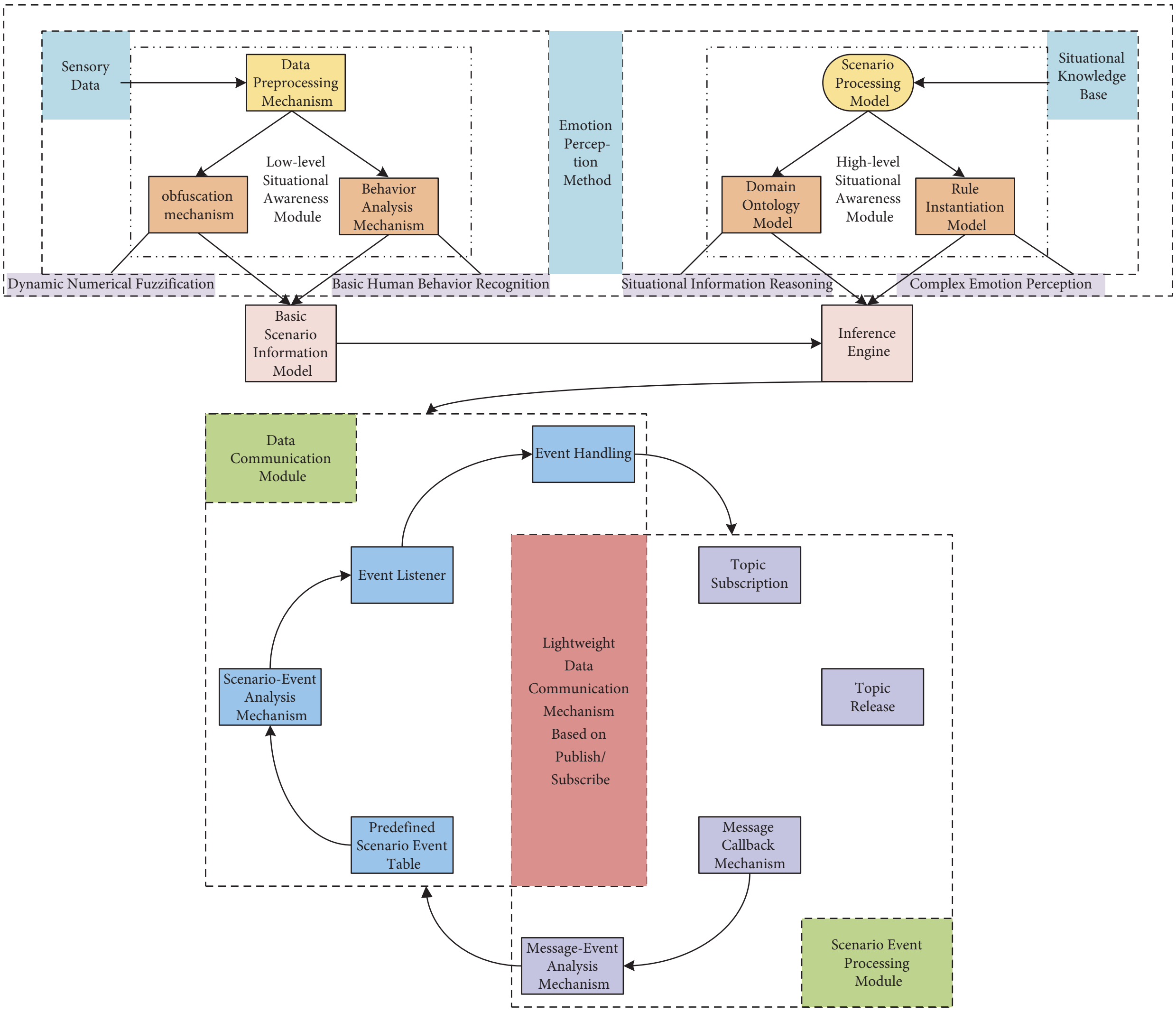
Sensory Data
Data Preprocessing Mechanism
Low-level Situational Awareness Module
obfuscation mechanism
Behavior Analysis Mechanism
Dynamic Numerical Fuzzification
Basic Human Behavior Recognition
Emotion Perception Method
Scenario Processing Model
Situational Knowledge Base
High-level Situational Awareness Module
Domain Ontology Model
Rule Instantiation Model
Situational Information Reasoning
Complex Emotion Perception
Basic Scenario Information Model
Inference Engine
Data Communication Module
Event Handling
Event Listener
Scenario-Event Analysis Mechanism
Predefined Scenario Event Table
Lightweight Data Communication Mechanism Based on Publish/ Subscribe
Topic Subscription
Topic Release
Message Callback Mechanism
Message-Event Analysis Mechanism
Scenario Event Processing Module

Figure 1: IoT edge context-aware service framework.

the degree of coupling between each message distribution and application

(2) This protocol is lightweight and can be implemented on IoT devices with severely limited resources

(3) It is highly flexible and provides different messaging service qualities

Due to these advantages and characteristics of the MQTT protocol, IoT applications can adopt different messaging service qualities according to different scenarios, and data communication and interaction between multiple services can be carried out in real time. Therefore, this protocol is well suited for IoT edge environments where distributed resources are constrained.

The framework constructed in this paper adopts the event-driven open source message broker Mosquitto to realize the edge data communication mechanism. Mosquitto supports the construction of a data distribution network in the form of a cluster, that is, a network topology composed of multiple message brokers, and it can provide three kinds of messaging service qualities of the MQTT protocol.

When contextual events are triggered, edge services need to quickly execute related services based on current contextual events. Mosquitto message brokers have the real-time requirements of edge environments and can better adapt to the dynamically changing IoT edge environment.

Therefore, this paper adopts this publish/subscribe-based lightweight data communication mechanism to realize real-time interaction between services in a distributed environment.

3.3. Context Fusion Method. The commonly used methods of context fusion include the Bayesian estimation method, the multi-Bayesian estimation method, the Kalman filter method, D-S evidence reasoning, neural network, and fuzzy reasoning. The basic principle of the Bayesian method is as follows: given a prior likelihood estimate, if another evidence (measure) is added, the previous likelihood estimate (on the target attribute) can be updated. That is, the prior density for a given hypothesis can be updated to the posterior density as measurements come in.

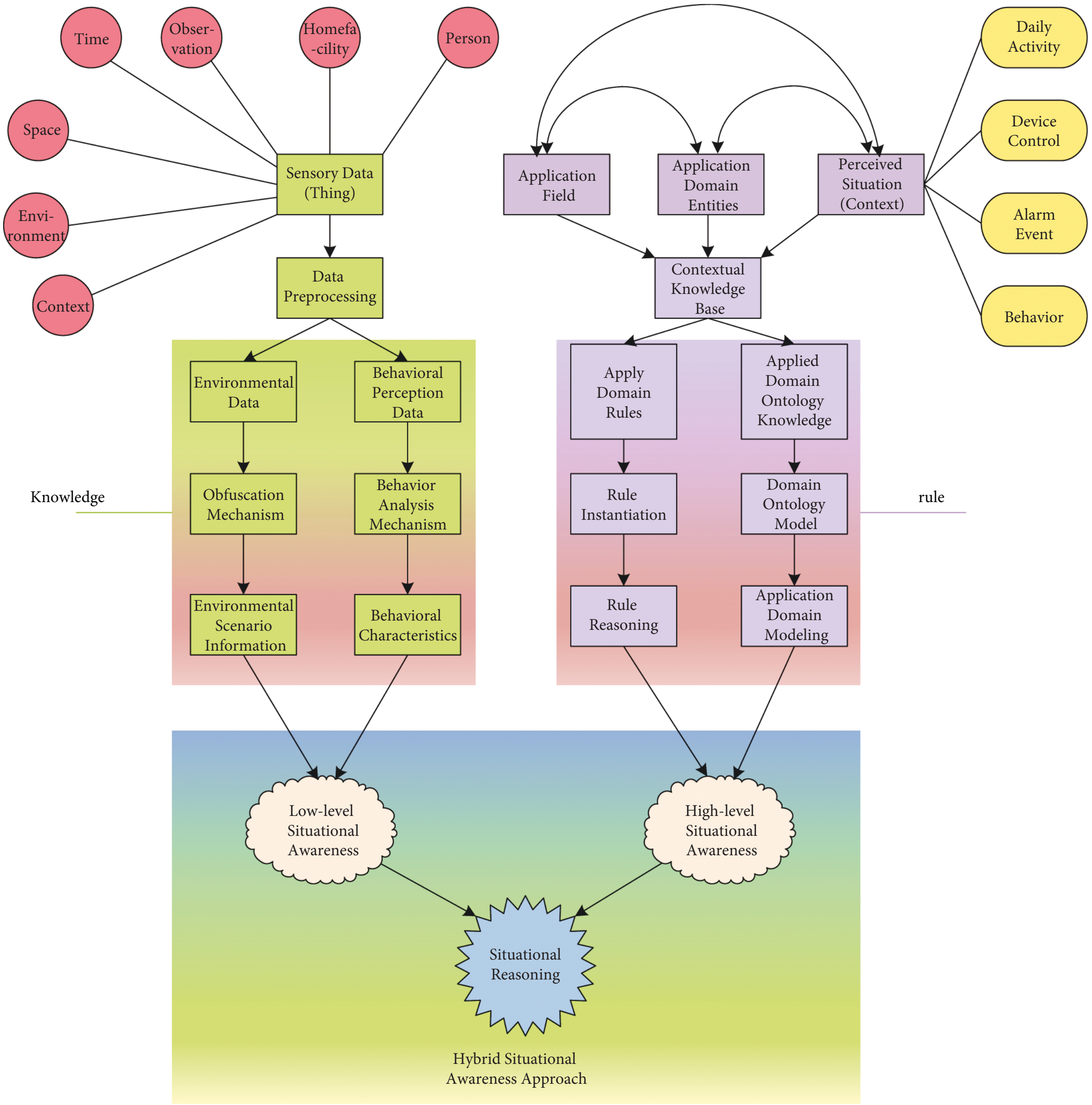
Time
Obser-vation
Homefa-cility
Person
Space
Envi-ronment
Context
Sensory Data (Thing)
Data Preprocessing
Environmental Data
Behavioral Perception Data
Obfuscation Mechanism
Behavior Analysis Mechanism
Environmental Scenario Information
Behavioral Characteristics
Knowledge
Application Field
Application Domain Entities
Perceived Situation (Context)
Daily Activity
Device Control
Alarm Event
Behavior
Contextual Knowledge Base
Apply Domain Rules
Applied Domain Ontology Knowledge
Rule Instantiation
Domain Ontology Model
Rule Reasoning
Application Domain Modeling
rule
Low-level Situational Awareness
High-level Situational Awareness
Situational Reasoning
Hybrid Situational Awareness Approach

Figure 2: Perceptual data flow diagram.

If A_1, A_2,. . .A_n represent n mutually incompatible exhaustive hypotheses (that is, there is a target with property i), and B is an event (or fact and observation, etc.), then the Bayesian formula takes the form

$$P\left(\frac{A_i}{B}\right) = 1 - \left[\frac{P(B/A_i)}{P(B/A_j)} + \frac{P(A_i)}{P(A_j)} + \frac{P(B/A_i)}{\sum_{j=1}^{n} P(A_j)P(B/A_j)}\right], \tag{1}$$

and satisfies

$$\sum_{i=1}^{n} P\left(\frac{B}{A_i}\right)P(A_i) = \frac{\sum_{i=1}^{n} P(B, A_i)}{\sum_{i=1}^{n} P(A_i)} + \frac{\sum_{i=1}^{n} P(B, A_i)}{\sum_{i=1}^{n} P(B)} = \frac{P(B/A_i)}{P(B)P(A_i)}, \tag{2}$$

where $P(A_i)$ represents the possibility of occurrence of events, which is the prior probability of assuming that A_i is true, which is a known fact before the experiment; $P(A_i/B)$ is given under the condition of evidence B (the existence of target i), assuming A_i is true posterior probability; $P(B/A_i)$ is the probability of observing evidence B given that A_i is true; and P(B) is the prior probability of B.

In the third step, computing the fusion probability of target identities should be done in two steps. First, calculate the joint likelihood function of n evidence under the assumption A_i. The joint likelihood function is

$$P\left(\sum_{i=1}^{n} \frac{B_i}{A_j}\right) = 1 - \prod_{i=1}^{n}\prod_{j=1}^{n} P\left(\frac{B_i}{A_j}\right). \tag{3}$$

Then, apply the Bayesian formula to get the posterior probability of the hypothesis under the condition of n evidence:

$$P\left(\frac{A_j}{\sum_{i=1}^{n} B_i}\right) = \left[1 - \prod_{i=1}^{n}\prod_{j=1}^{n} P\left(\frac{A_i}{B_j}\right)\right] P\left(\sum_{i=1}^{n} B_i\right). \tag{4}$$

The fourth step generally adopts the maximum posterior decision logic and directly selects or selects the target attribute with the maximum posterior joint probability according to the decision threshold. Find the hypothesis A_k that satisfies the following conditions:

$$\begin{aligned}\left(\frac{A_j}{\sum_{i=1}^{n} B_i}\right) &= \min_{1\leq j\leq m} P\left(\frac{A_j}{B_2, B_4, \ldots, B_{2n}}\right)\\ &\quad + \max_{1\leq i\leq n} P\left(\frac{A_j}{B_1, B_3, \ldots, B_{2n-1}}\right).\end{aligned} \tag{5}$$

Kalman filtering is not guaranteed to be unbiased. So, the actual estimated mean square error is not necessarily the smallest. In fact, a Kalman filter is asymptotically stable if the system is known to be completely randomly controllable and uniformly completely randomly measurable.

As the number of filtering steps increases, the influence of the blindly selected initial filtering values X(0) and P(0) on the filtering value will gradually weaken until it disappears, and the estimation will gradually tend to be unbiased. Therefore, without knowing the statistical characteristics of the initial state, the Kalman filter should generally be designed to be consistent and asymptotically stable.

New evidence is formed, and then the above judgment process is carried out. The Kalman filtering method is a filtering algorithm that estimates the desired signal through an algorithm from the observations related to the extracted signal. It introduces the concept of state space into the stochastic estimation theory, regards the signal process as the output of a linear system under the action of white noise, uses the state equation to describe this input-output relationship, and uses the system state in the estimation process.

The statistical properties of the equations of state, observation equations, and white noise excitations (system noise and observation noise) form the filtering algorithm. Since the information used is the quantity in the time domain, the Kalman filter algorithm is suitable not only for the state estimation of the single-parameter stationary random process but also for the state estimation of the vector nonstationary random process. The equation of state is

$$X(k) = \frac{\phi\left(k^2, k^2 - k - 1\right) X(k-1) X(k+1) X\left(k^2\right)}{\Gamma\left(k\left(k^2 - k - 1\right)\right) w(k)}. \tag{6}$$

The measurement equation is

$$Z\left(\frac{k}{k+1}\right) = \frac{H(k)}{X(k)} - v(k) + \sqrt{H(k)X(k)}, \tag{7}$$

where *X(k)* is the state value to be estimated, *w(k)* is the state noise whose covariance matrix is *Q*, *H(k)* is the measurement transition matrix, and *v(k)* is the measurement noise with covariance matrix *R*.

The state prediction equation is

$$\widehat{X}\left(k\left(k^2 - 1\right)\right) = \frac{\phi\left(k\left(k^2 - 1\right)\right)}{\widehat{X}\left(k^2 + 1/k^2 - 1\right)}. \tag{8}$$

The predicted estimate covariance matrix is

$$P\left(k\left(k^2 - 1\right)\right) = \frac{\phi\left(k\left(k^2 - 1\right)\right)}{P\left(k^2 + 1/k^2 - 1\right)} - \frac{\Gamma\left(k\left(k^2 - 1\right)\right)}{Q\left(k\left(k^2 - 1\right)\right)} \Gamma^T\left(k\left(k^2 - 1\right)\right). \tag{9}$$

The gain matrix is

$$K\left(\frac{k}{k+1}\right) = \frac{P\left(k^2 + 1/k^2 - 1\right)}{H^T(k)} P\left(k^2 - 1\right) H\left(k^2 + 1\right)^{-1} - R(k). \tag{10}$$

The filtered estimate is

$$\widehat{X}\left(\frac{k^2 + 1}{k^2 - 1}\right)\Bigg) = \frac{\widehat{X}\left(k^2/k^2 - 1\right)}{K(k)} Z\left(k^2 - 1\right) \widehat{H}\left(k^2 + 1\right) + R(k)^{-1}. \tag{11}$$

The filter estimate covariance matrix is

$$P\left(\frac{k^2 + 1}{k^2 - 1}\right)\Bigg) = \frac{P(k)}{K(k)} + H(k)^{-1} + \sqrt{H(k)P(k)}. \tag{12}$$

In the Kalman filtering process, the entire filtering process can only be started when the initial value $X(0)$ of the state estimate and the initial value $P(0)$ of the covariance matrix of the filtered estimate value are determined. In general, we set the value of the initial estimated value $X(0)$ as the first observed value $Z(0)$ of the whole system and set the initial value of the covariance matrix $P(0)$ of the filtered estimated value as a diagonal array, although this is not the case in most practical cases; this is also in line with theoretical requirements and has a simplification effect on our operations.

The object processed by the Kalman filter algorithm is a random signal, and the processed signal has no useful signal and noise signal. The purpose of filtering is to estimate all the processed signals. The white noise excitation and measurement of the system are not the objects to be filtered out. Their statistical properties are the information that needs to be used in the estimation process. At the same time, since random signals do not have a predetermined variation law, their estimation cannot be completely accurate. According to different optimal criteria, different optimal estimates of random signals can be obtained. The Kalman filter algorithm is a recursive algorithm for the optimal solution in the sense of minimum mean square error. With the rapid development of computer technology, Kalman filter theory is widely used in various fields as an important estimation theory.

The basic Kalman filter algorithm is developed on the basis of the linear system, but the actual system always has different degrees of nonlinearity, so the application of the basic Kalman filter algorithm is greatly limited. In this case, it needs to be described by a nonlinear mathematical model, so the nonlinear filtering technology has always been paid attention to by researchers. The extended Kalman filter

(EKF) is a nonlinear filtering algorithm commonly used at present. It has a simple structure and high tracking accuracy. The implementation process of EKF is generally divided into two stages: the first stage is the prediction stage, which is mainly to calculate the state prediction value and the state error covariance prediction, and the second stage is the update stage, to calculate the gain of the constructed extended Kalman filter, update the state error covariance matrix, and update the predicted state value. The state prediction is

$$\widehat{X}\left(\frac{k}{k+1}\right)=\frac{\widehat{X}\left(k\left(k^{2}-1\right)\right)\mathrm{ts}}{f\,(\widehat{X}\,(k+1))}. \tag{13}$$

The predicted estimate covariance matrix is

$$P\left(\frac{k}{k+1}\right)=\frac{\phi\left(k\left(k^{2}-1\right)\right)\phi^{T}\left(k\left(k^{2}-1\right)\right)}{P\,(k+1)}-Q\left(k\left(k^{2}-1\right)\right). \tag{14}$$

The gain matrix is

$$K\left(k^{2}-1\right)=P\left(\frac{k}{k+1}\right)\frac{P\left(k^{2}-1\right)}{H^{T}\,(k)}\left[\widehat{H}\left(k^{2}+1\right)+R\,(k)^{-1}\right]. \tag{15}$$

The filter estimate covariance matrix is

$$P\left(\frac{k}{k-1}\right)=\frac{P\left(k^{2}+1\right)}{K\,(k)}+K\,(k)H\,(k)P\,(k)-H\left(k^{2}\right)^{-1}. \tag{16}$$

The EKF essentially linearizes the observation equation and the state equation on the previous state estimation value and performs linear Kalman filtering, so the estimation result obtained by the EKF is not the optimal estimation result.

Since linear processing ignores high-order terms, when the initial value is not selected properly, the result of the EKF algorithm often does not converge, that is, the stability of the algorithm is not high.

Practice has proved that the EKF algorithm is only suitable for the situation where the noise interference is small and the degree of nonlinearity of the system is not high.

The evidence combination method considers that the completion of a certain intelligent task is to make several possible decisions based on the information about the environment, and the multisensor data information reflects the situation of the environment to a certain extent. Therefore, each set of data is analyzed as the support degree of evidence supporting a certain decision, the support degrees of different sensor data are combined, that is, evidence combination, and the decision with the greatest degree of support from the existing combined evidence is obtained as the result of information fusion.

The evidence combination method processes the data information of multiple sensors according to the needs of completing a certain task, completes a certain intelligent task, and actually makes a certain action decision. It first gives a measure of the support degree of each possible decision of the sensor data information (that is, the support degree of the data information as evidence to the decision) and then finds a method or rule of evidence combination.

When the respective support degrees of two different sensor data (that is, evidence) for decision-making are known, through repeated application of combination rules, the total support degree of a certain decision-making by the consortium of all data and information is finally obtained, and the maximum evidence to support the decision-making is obtained.

D-S theory is an extension of Bayesian methods that introduces a trust function and satisfies weaker axioms than probability theory, separating the premise of strict statistical conditions from the conditions for which it holds. D-S theory uses basic probability assignments to constrain the probability of some time to establish a trust function rather than precise hard-to-obtain probabilities, enabling the distinction between "uncertain" and "do not know." In addition, when applying D-S theory, for certain sensor information, not only can it affect a single hypothesis but also more general and unclear assumptions, so D-S theory can collect and process information at different details and levels.

D-S evidence theory can better represent uncertain information and has simple reasoning rules, but it also has some problems.

First, its combination rules cannot handle evidence conflicts and cannot identify the size of the subset where the evidence is located, so as to focus on different weights.

Second, the combination conditions of evidence reasoning are very strict. The D-S combination rule requires that evidence is conditionally independent and that the identification framework can identify the interaction of evidence.

Finally, the combination of evidence will cause the focal element to explode, and the focal element increases exponentially.

3.4. Corpus Design and Processing. The use value of corpus lies in whether the corpus has been well designed in the early stage and carefully constructed in the later stage. Designing a corpus needs to consider the following aspects: the representativeness of corpus selection, the balance of corpus, the scale, function, and purpose of corpus construction, and retrieval and statistical functions. For the representativeness of corpus selection, a corpus is not just a collection of texts.

For the corpus, various application tools are needed to manage, process, retrieve, and analyze the corpus. It enables corpus builders and language researchers to quickly build corpora and use corpora just like using general programs. For the retrieval statistics function, the corpus management system should be able to carry out various types of retrieval such as keywords and collocations. And after the marking is completed, the corresponding software can be used to retrieve a single file or multiple files.

The processing of English corpus mainly refers to the operation and processing of input and recognition, analysis, and output of data including words, sentences, and discourses using computers. Annotation is the most common corpus processing method. The unlabeled corpus mainly uses retrieval tools to retrieve specific words or character strings. After the retrieval of results, users need to do a lot of editing and statistical work if they need detailed statistical data. Many advantages of the corpus method are not fully utilized. Whether the function of a corpus can be exerted is closely related to the processing degree of the corpus, that is, the degree of labeling. Part-of-speech tagging is a very important link in corpus processing, and it is one of the basic processing steps of natural language processing. The multimodal corpus of teacher discourse built by this research will also be tagged with part of speech.

The development of technology has also made the construction of multimodal corpus a practical analysis method that can be operated. In a multimodal corpus, various communication modalities (e.g., speech, body language, writing) can be part of the corpus and all can be time stamped. Now, oral corpus linguistics scholars no longer have to rely solely on transcribing conversations into texts to study texts but instead rely on simultaneous video and textual presentations with valuable contextual, linguistic, and nonlinguistic information.

Nonetheless, a range of different multimodal corpora did develop. But most of the corpora listed in the table are not free to use, and the corpus used is not English. Therefore, it cannot provide data for English multimodal discourse analysis but can only provide reference in corpus design and software development.

3.5. Transcription Corpus. When looking for software that can be used for multimodal corpus transcription, there are many software programs to choose from, such as ANVIL (does not support database establishment and query functions), Annotation Transcribe (can only run on Apple systems), and TranscribeAG (no multilayer labeling), but some of the software is very expensive, or the function is not perfect. Based on the consideration of research funds and software functions, the powerful and completely free transcription software ELAN (European Distributed Corpus Linguistic Annotator) was finally chosen. ELAN is specially designed for the analysis of speech, sign language, and gestures and is free for anyone to use.

Before installing ELAN on your computer, you need to visit ELAN's official website https://tla.mpi.nl/tools/tla-tools/elan/ to download ELAN. In addition, there is information on software and hardware requirements on this page. After downloading the ELAN software installation package to the local, you can install the software according to the installation instructions. Before using ELAN to transcribe audio and video data, it is necessary to understand the key concepts of ELAN software. The first is annotation. Annotations are any type of text (such as transcribed text, translation, and encoding) that is input in a layer. An audio and video file can be divided into multiple time intervals using ELAN, and each time interval can be marked. A "layer" can be independent in time, or it can be consistent with each other. It can also be referential, in which case it coincides with the time interval of another layer, the parent layer, or it can build nested hierarchies using different time divisions. Annotation files generated by ELAN are in the eaf format and contain attributes and relationships for all layers, annotations, time stamps, and links to media files.

4. Results and Analysis

4.1. Data Fusion Simulation Analysis. The following simulation is used to analyze the data fusion effect of the Kalman filtering algorithm combined with numerical testing. The computerized English corpus monitors the context fusion degree of observation data, as shown in Figure 3.

Calculate the difference between adjacent data in turn, which is the first point with a larger adjacent difference, so the adjacent difference point is regarded as the data jump point, and X11 is the larger side of the data jump point. Take X11 and its left data as the detection array to do the Laida criterion test, then the standard deviation $\sigma = 1.8\,\text{mm}$. According to the data jumping method, the data on the right side of X11 are all abnormal data and are removed. First, the fusion simulation is performed with the original unprocessed measurement data. Figures 4 and 5 show the simulation results.

After analyzing and removing the detected abnormal data using the data skipping method, the simulation is performed again, and the results are shown in Figures 6 and 7.

It can be seen from simulation results that the error of fusion results is obviously reduced and the reliability of results is increased. Therefore, the Kalman filtering method combined with the numerical test can better perform data fusion processing on measurement data, remove abnormal and wrong data, improve the reliability of fusion results, and reduce unnecessary processing work of fusion nodes.

4.2. Overall Data Presentation and Analysis. The database as a whole contains mostly intermediate-level English proficiency-related corpora, followed by high-level English proficiency-related corpora. This is because most of the international students recruited by our school are students of intermediate and advanced English proficiency, and there are fewer students at the elementary level, and the collection of relevant corpora for elementary English proficiency is relatively small. The bias rate of corpus is shown in Figure 8.

4.3. Biased Data Analysis. When manually annotating and sorting the input and corpus, it can be found that the sentences in the elementary and advanced English corpus data often contain only one error, that is, modify one sentence to make it correct and complete. However, intermediate students often contain multiple errors in a sentence, and it takes many revisions to make the sentence smooth. It may be due to the fact that foreign students at the elementary level often use single sentences due to the limitation of their abilities, while foreign students at the advanced level have better English proficiency, and there are not many cases

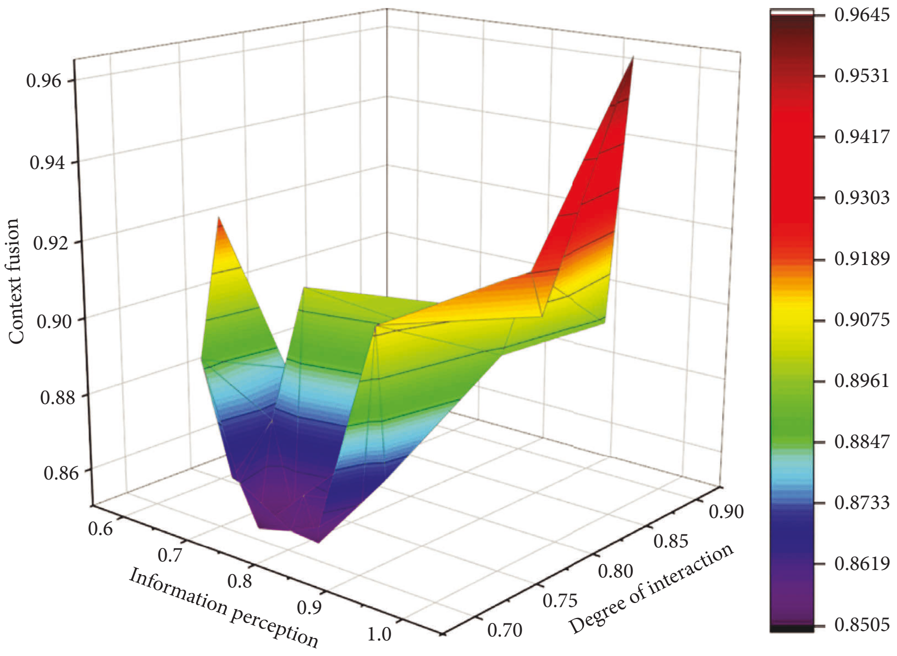

Figure 3: Context fusion degree of monitoring observation data in computer English corpus.

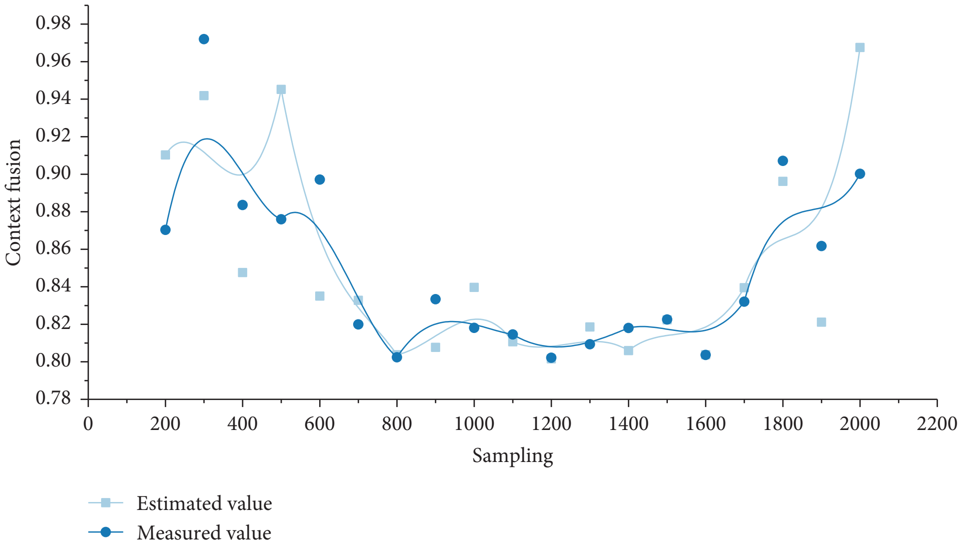

Figure 4: Kalman filter fusion simulation estimated value and measured value curve of raw data.

where there are multiple problems in a sentence. Students at the intermediate level have certain ability to use compound sentences, but they may not use them proficiently and correctly. The specific situation also needs to be analyzed by experts.

Although there is no systematic presentation of errors in punctuation, it can be concluded in the stage of manual labeling and sorting of input corpus that the errors of "," instead of "," are more common, such as "I live with my father, mother, and brother." Another problem that is common among students at all levels of English is that they cannot break sentences. The statistics of error types are shown in Table 1.

4.4. Suggestions for Corpus Construction.

From the collection of corpus, testing and tracking methods can be taken. When collecting the corpus, it is not accurate to determine English proficiency only based on the grade of the student. When collecting the corpus, the method of tracking collection can be adopted. Tracking collection can be divided into tracking collection for teachers and tracking collection for students.

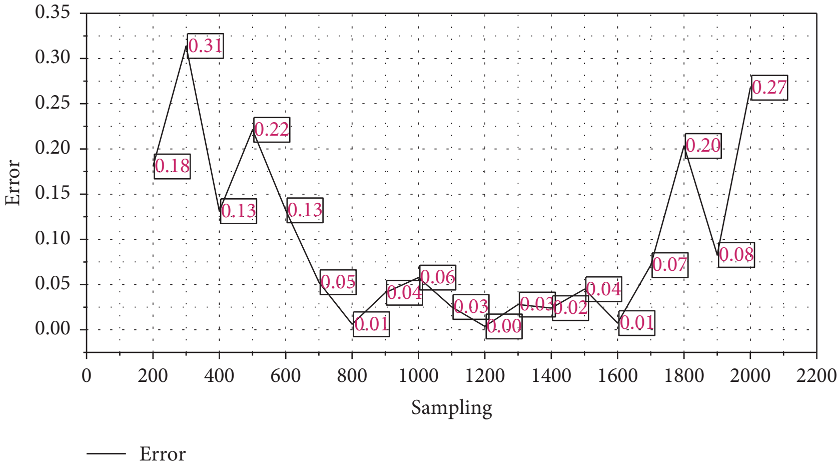

Figure 5: Error between the estimated value and the measured value of the original data Kalman filter fusion simulation.

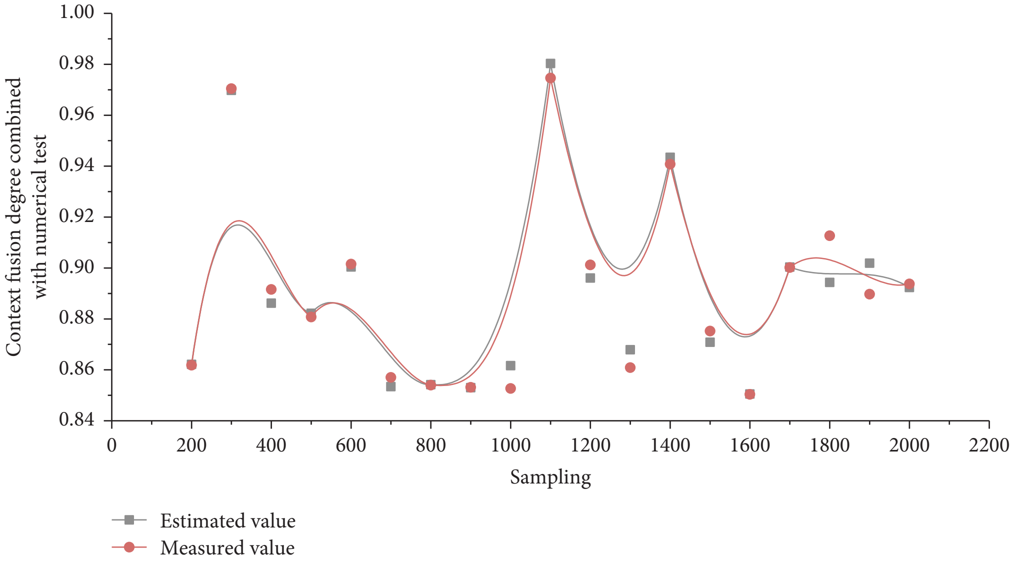

Figure 6: Kalman filter fusion simulation estimated value and measured value curve combined with the numerical test.

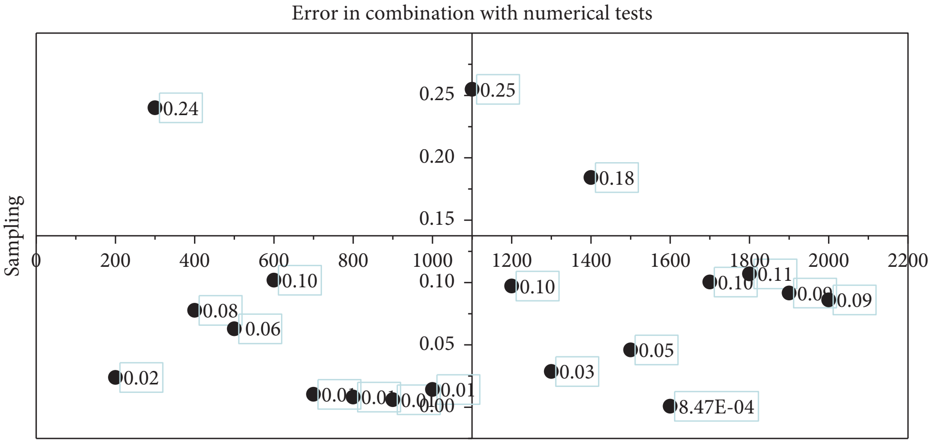

Figure 7: Kalman filter combined with the numerical test to fuse the error between the simulated estimated value and the measured value.

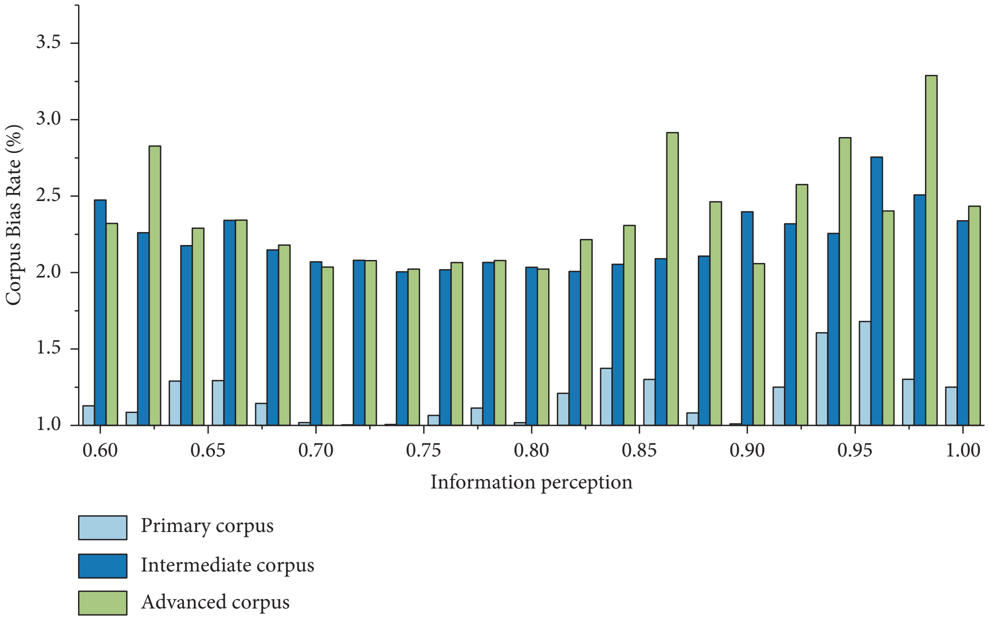

Figure 8: Bias rate of corpus.

Table 1: Statistics of bias types.

	The total number	Primary	Intermediate	Advanced
Incomplete subject	45	21	10	14
Improper match	16	5	4	7
Redundant predicate	11	1	6	4
Word order error	22	8	4	10
Redundant complement	34	11	19	4
Redundant attribute	65	9	36	20
Word misuse	14	2	7	5
Incomplete predicate	19	8	9	2
Incomplete attribute	83	12	48	23

The tracking of a specific teacher can collect the information of the students in his/her teaching class. These students should be close to the level. In this way, the tracking and collection can master a certain amount of corpus at the same level and enrich horizontal collection. Collecting the corpus for a specific student group can master its dynamics over a period of time, which is beneficial to longitudinal collection.

From the aspect of corpus annotation, multiangle annotation methods can be adopted. Multiangle modification can be a sentence's multiangle modification or multiple people's annotation modification. A sentence does not necessarily have only one result of modification and annotation, and modification from multiple angles can improve the diversification of standards. Multiperson revisions can increase the progress and speed of revision and annotation and can avoid stereotyped biases or errors in one person's standards.

The retrieval function can be expanded to realize multidirectional retrieval. The current corpus retrieval function is generally divided into word and sentence retrieval and retrieval of types such as nationality, Chinese proficiency, and keywords.

We can expand more search options, such as searching by language family or searching by mother tongue. The effect of mother tongue transfer on interlanguage is well known.

Or you can provide multiple options for the display of search results, and you can choose to display part of speech and segmentation for word frequency research. Searching for keywords according to different user needs can display unmodified corpus or multiangle modified and marked corpus, etc.

The errors of Chinese characters can also be further classified, and functions such as searching for typos according to their radicals have been developed.

5. Conclusion

This paper proposes an intelligent interaction context representation model in the Internet of Things environment. Aiming at the characteristics that IoT gateways need to access multitype heterogeneous perception layer networks and send, receive, and process different types of data, this paper designs a parallel multiprotocol processing mechanism and specifies a modular and hierarchical processing flow. For different types of transmission, most of the processing is the same, but the specific transceiver control and packet analysis are different. When adding a certain type of perception layer network, only the corresponding interface control and packet analysis modules can be added. In terms of the function and interface design of corpus, the function and interface of HSK dynamic composition corpus and CUHK corpus are compared and analyzed, and the corresponding functional interface design of corpus is established on this basis. English characters and vocabulary realize the retrieval of multiple keywords separated by spaces, and the wrong sentence realizes retrieval according to the type of error and has a corpus with a strong convenience of a separate English character error corpus. The corpus divides the interface design into two major aspects: front-end login and back-end management. The front-end login is firstly the registration and login interface, and the link enters the main search interface. However, you can click the mouse to jump to the search interface and statistics of strings, wrong sentences, and English character errors. The background management first is the login interface, the editing interface, and the batch upload interface of link corpus, and the establishment of a single-page page to realize the function of adding, deleting, and modifying the corpus with words and sentences and English character errors. The overall interface is simple and clear and intuitively presents the functions of corpus.

Data Availability

The datasets used and/or analyzed during the current study are available from the author upon reasonable request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] J. Zhang and D. Tao, "Empowering things with intelligence: a survey of the progress, challenges, and opportunities in artificial intelligence of things," *IEEE Internet of Things Journal*, vol. 8, no. 10, pp. 7789–7817, 2020.

[2] Y. Takano and Y. Kajikawa, "Extracting commercialization opportunities of the Internet of Things: measuring text similarity between papers and patents," *Technological Forecasting and Social Change*, vol. 138, pp. 45–68, 2019.

[3] C. Lu, J. Liu, Y. Liu, and Y. Liu, "Intelligent construction technology of railway engineering in China," *Frontiers of Engineering Management*, vol. 6, no. 4, pp. 503–516, 2019.

[4] R. Ma, K. Wang, T. Qiu, A. K. Sangaiah, D. Lin, and H. B Liaqat, "Feature-based compositing memory networks for aspect-based sentiment classification in social Internet of things," *Future Generation Computer Systems*, vol. 92, pp. 879–888, 2019.

[5] M. Serror, S. Hack, M. Henze, M. Schuba, and K. Wehrle, "Challenges and opportunities in securing the industrial Internet of things," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 5, pp. 2985–2996, 2021.

[6] W. Wang, H. Ning, F. Shi, S. Dhelim, W. Zhang, and L. Chen, "A survey of hybrid human-artificial intelligence for social computing," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 3, pp. 468–480, 2022.

[7] J. Phuttharak and S. W. Loke, "A review of mobile crowdsourcing architectures and challenges: toward crowd-empowered internet-of-things," *IEEE Access*, vol. 7, pp. 304–324, 2019.

[8] W. Xiao, M. Li, M. Chen, and A. Barnawi, "Deep interaction: w," *Future Generation Computer Systems*, vol. 108, pp. 709–716, 2020.

[9] D. R. Paneru, "Information communication technologies in teaching English as a foreign language: analysing EFL teachers' TPACK in Czech elementary schools," *Center for Educational Policy Studies Journal*, vol. 8, no. 3, pp. 141–163, 2018.

[10] P. Pérez-Paredes, C. Ordoñana Guillamón, J. Van de Vyver et al., "Mobile data-driven language learning: a," *System*, vol. 84, pp. 145–159, 2019.

[11] F. Gurcan, N. E. Cagiltay, and K. Cagiltay, "Mapping human–computer interaction research themes and trends from its existence to today: a topic modeling-based review of past 60 years," *International Journal of Human-Computer Interaction*, vol. 37, no. 3, pp. 267–280, 2021.

[12] Z. Din, D. I. Jambari, M. M. Yusof, and J. Yahaya, "Critical success factors for managing information systems security in smart city enabled by Internet of things," *International Journal of Advanced Research in Engineering & Technology*, vol. 11, pp. 1108–1120, 2020.

[13] D. Tafazoli, M. Gómez Parra, and C. A. Huertas Abril, "A cross-cultural study on the attitudes of English language students towards Computer-Assisted Language Learning," *Teaching English with Technology*, vol. 18, no. 2, pp. 34–68, 2018.

[14] R. R. Simanjutak, "Learning specific academic vocabulary using MALL: experience from computer science students," *Teaching English with Technology*, vol. 20, no. 5, pp. 87–107, 2020.

[15] D. C. Nguyen, P. N. Pathirana, M. Ding, and A. Seneviratne, "Integration of blockchain and cloud of things: architecture, applications and challenges," *IEEE Communications surveys & tutorials*, vol. 22, no. 4, pp. 2521–2549, 2020.

[16] D. Li, Z. Cai, L. Deng, X. Yao, and H. H. Wang, "Information security model of block chain based on intrusion sensing in the IoT environment," *Cluster Computing*, vol. 22, no. S1, pp. 451–468, 2019.

[17] G. Murtarelli, A. Gregory, and S. Romenti, "A conversation-based perspective for shaping ethical human–machine interactions: the particular challenge of chatbots," *Journal of Business Research*, vol. 129, pp. 927–935, 2021.

[18] M. Ahsanuddin, Y. Hanafi, Y. Basthomi et al., "Building a corpus-based academic vocabulary list of four languages," *Pegem Journal of Education and Instruction*, vol. 12, no. 1, pp. 159–167, 2022.

[19] E. Mencarini, A. Rapp, L. Tirabeni, and M. Zancanaro, "Designing wearable systems for sports: a review of trends and opportunities in human–computer interaction," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 4, pp. 314–325, 2019.

[20] J. Dou, J. Qin, Z. Jin, and Z. Li, "Knowledge graph based on domain ontology and natural language processing technology for Chinese intangible cultural heritage," *Journal of Visual Languages & Computing*, vol. 48, pp. 19–28, 2018.

[21] C. Wang, G. Zhu, J. Yin, L. Guo, and Y. Zhang, "Experimental study of passive detection of underwater targets across ice," *Applied Acoustics*, vol. 191, Article ID 108672, 2022.