Design of a Natural Language Information Extraction and Proofreading System Based on AI Algorithms

Wenbin Cai, Yanqing Chen[*]

International Business College, South China Normal University, Foshan, Guangdong, China
[*]Corresponding author

Abstract: *Traditional manual proofreading methods can no longer meet the demands of efficient dissemination of data and information under the development of information technology. Therefore, a natural language information processing system for the legal consulting market in new media environments based on AI algorithms is proposed. The system can automatically search for important page information and text, extract text, generate summaries, and proofread and review information, then send the proofread information to searchers, enabling precise pre-screening of all information and saving reading effort and cost. Through experimental testing of the system's algorithmic functions, the system's operational results meet expectations.*

Keywords: *AI algorithms; natural language; information extraction; information proofreading*

1. Introduction

Against the background of modern global integration, the popularization of digital infrastructure has increased the speed of information generation and dissemination. However, the lack of a unified data screening mechanism has resulted in massive volumes of content. In addition, news platforms and social media rely on algorithms to prioritize highly interactive content, which amplifies fragmented and emotional information, making the problems of information overload and scarcity of effective information increasingly prominent. The development of artificial intelligence technology applies natural language processing techniques to reasonably address such problems, processing millions of web pages within 1 second and extracting the information required by information search users from new media, news information, and forum web pages based on the natural linguistic characteristics of information [1]. Through information integration, noise processing, redundancy removal, and other operations, unstructured information is transformed into structured information. H. P. Lnhn [2] first proposed information extraction technology in 1958, summarizing information by extracting features such as phrase and sentence word frequencies from textual information. Subsequently, researchers proposed the concept of cue words on this basis, determining the weight of sample titles and main text sentences, and designed machine translation systems based on AI technology to recognize different forms of characters. In research on natural language information extraction, both domestic and international studies have only completed information extraction without proofreading the information. Accordingly, this paper studies the extraction and proofreading of search reading page information based on artificial intelligence algorithms, completing page pre-screening functions and saving effort and time in information extraction.

2. Architecture of the Natural Language Information Extraction and Proofreading System

To enhance the proofreading effect after extracting natural language information, this paper develops the intelligent proofreading prediction function module shown in Figure 1. The system is composed of units such as the work module, search module, proofreading module, and user module, thereby realizing intelligent extraction and proofreading of natural language.

Through analysis of the working framework of the natural language information extraction and proofreading system, the system is essentially an analysis and processing system for news, market, and media content information. By mining useful information from news and intelligently proofreading the information, data with higher readability and accuracy are obtained. During intelligent proofreading of public information, a large amount of operational data is generated and stored in the log unit. These data can provide backend engineers with real-time monitoring of system operations, thereby improving the accuracy of system proofreading [3].

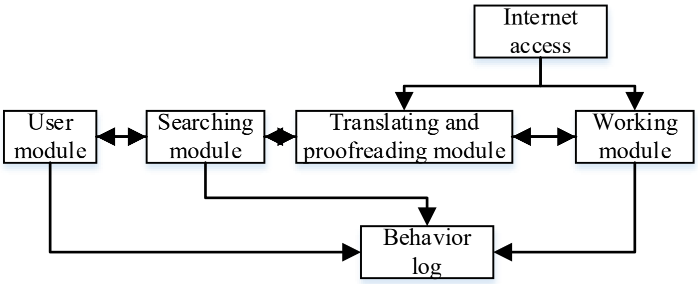

Figure 1. Functional framework of the extraction and proofreading system

3. Design of the Natural Language Information Extraction and Proofreading System

3.1 Hardware Selection of the System

To ensure the accuracy of natural language information extraction and recognition, this paper configures an optocoupler with an operating frequency set to 125–335 Hz, combined with sensing equipment to form a cyclic information collection and processing structure. The latch selected is the 54LS273 model, and information is identified through a microcontroller. Table 1 presents the indicators and parameters of hardware control. In addition, the SM2246XT main control chip is added to the system hardware to realize the function of directional information guidance and control.

Table 1. Indicators and Parameters of Hardware Control

Hardware indicator	Standard parameter value	Controllable range
Voltage/V	27	32
Output power conversion ratio	3.25	4.15
Output current/mA	500	550
Multistage voltage regulation difference/V	3.1	2.5

3.2 Software Functions of the System

3.2.1 Data Collection and Preprocessing

In the system design process, the collection and processing of high-quality corpora is one of the most critical steps. In this paper, the corpus is constructed using legal news texts, legal statute databases, and legal case repositories to create a large-scale academic corpus. Due to the large volume of corpus data, a distributed streaming preprocessing pipeline is designed, utilizing the Spark Streaming framework to perform parallel processing of the data, including data segmentation, cleaning, syntactic analysis, and part-of-speech processing. Operations are carried out in a logical sequence, improving the effectiveness of data collection and preprocessing. The cleaning module is used to filter low-quality samples, ensuring the readability of the corpus.

In terms of technology, the parsing accuracy is improved through the syntactic parsing capability of the Stanford CoreNLP tool. A metadata management mechanism is established to evaluate the results of data preprocessing in real time. A reasonable neural network semantic model is constructed to analyze the semantics and syntax of texts. In this paper, a 24-layer mutual attention network combined with positional encoding is used to enhance the modeling capability for long texts. In addition, an upgraded pre-training framework is adopted to increase the complexity of the corpus. The grammatical representation formula is expressed as:

$$J_\theta = \max_\theta \sum_{t-1}^{T} \log p_\theta(\omega_t / \omega < t) \tag{1}$$

In Equation (1), the text sequence is represented by ω_t, and θ denotes the parameters of the language model.

In addition, considering the need for manual adjustment, the model design in this paper can identify erroneous content in specific types of texts and derive a fine-tunable language model, thereby improving the error detection efficiency of the system [4].

3.2.2 Clustering Analysis of Data Information

Data mining techniques are applied to perform clustering analysis on legal service user consultation content and legal topics on new media platforms, in order to identify common types of legal issues and the distribution of user demands. In data mining, when the same information data are mined, nodes need to be labeled and information allocated. Data mining is used to analyze the system database on the server side, listing information such as data warehouses, model time, and algorithms. Mining data are described using data tables and tree structures, which makes it possible to classify user value types among unknown attributes. For example, consultation clients can be classified into user demand types such as low demand, medium demand, high demand, and extremely high demand, and different management measures can be formulated for different client types. The clustering results of consultation client value are shown in Table 2. Based on the classification principles, consultation client management information is analyzed to support service decision-making driven by technology.

Table 2. Value Clustering Results of Consultation Clients

No.	User demand type	Legal complexity	Consultation frequency
1	Low demand	Low	Low
2		Moderate	Low
3	Medium demand	Low	Moderate
4		High	Low
5		Moderate	Moderate
6	High demand	Moderate	High
7		High	Moderate
8	Extremely high demand	High	High
9		Moderate	High

3.2.3 New Word Recognition Function

(1) Mutual Information. New words refer to lexical items that are not included in the dictionary. They can be generated through multiple channels, including alphabetic words, newly coined terms, technical terminology, and named entities. The presence of new words may lead to “fragmented strings” in word segmentation, thereby causing problems in sentiment lexicons and Chinese word segmentation. Therefore, accurately identifying new words and standardizing them as linguistic units can improve the performance of automatic Chinese word segmentation. In Chinese sentences, each character is associated with others. If the degree of association is high, the probability of forming words between characters or between words and characters increases. The larger the mutual information value, the stronger the dependency between the two elements. The mutual information is calculated as follows:

$$MI(x,y)=\log\frac{p(x,y)}{p(x)p(y)} \tag{2}$$

In the formula, $p(x),p(y)$ and $p(x,y)$ denote characters or words, x,y represents the probabilities of the corresponding characters or words in the corpus.

(2) Left and Right Entropy. This refers to a statistical measure of the uncertainty of neighboring words or characters of a candidate word, calculated using information entropy. The greater the uncertainty, the more information the neighboring strings carry, which helps form meaningful words or phrases. The left and right adjacency entropies are defined as follows:

$$\begin{aligned} HR(x) &= -\sum_a p(a\,|\,x)\log p(a\,|\,x) \\ HR(x) &= -\sum_b p(b\,|\,x)\log p(b\,|\,x) \end{aligned} \tag{3}$$

In the formula, $p(a\,|\,x),p(b\,|\,x)$ represent the probabilities of the left and right adjacent characters of the two candidate words a and b for the word x, respectively.

(3) Combining Mutual Information with Left and Right Entropy.

Left and right entropy is used to filter new words. If the entropy value in Equation (4) exceeds the threshold, the candidate new word is retained; otherwise, the candidate new word is discarded.

$$\begin{aligned} &MI - HLR(word_i,...,word_j) \\ &= (MI(word_i,...,word_n) + HLR_{\min}) \times f(word_1,...,word_n) \end{aligned} \tag{4}$$

By comparing the new word recognition functions across different databases, it is shown that the new word recognition function of the algorithm proposed in this paper performs well [5].

3.2.4 Electronic Information Retrieval Database

By linking multiple information recognition nodes, an information monitoring and processing structure is formed. The natural language processing (NLP)–based query mechanism is designed to transform input into information retrieval statements usable by search engines, allowing the computation of retrieval efficiency for unidirectional information as follows:

$$M = \int \varpi^2 - \sqrt{(N+k)} \tag{5}$$

In the formula, ϖ represents the retrieval feature value, *M* is the unidirectional information retrieval ratio, *k* is the detection range, and *N* is the average retrieval value. The obtained retrieval ratio is used to transmit the collected information within the expected retrieval database. Combined with the AI algorithm, users are randomly selected, and the retrieval targets of language information are extracted, allowing the calculation of the detection feature values as follows:

$$Y = o - \frac{v^2 + t}{S} \tag{6}$$

In the formula, *o* represents the retrieval preference value, *Y* the feature value, *v* the satisfaction level, *S* the retrieval range, and *t* the storage quantity. Based on the calculated values in Equation (6), information is classified and stored in respective blocks, thereby improving the retrieval database. The data retrieval algorithm is as follows:

```
static void RetrieveMultipleResults(SqlConnection connection)
{
    using (connection)
    {
        SqlCommand command = new(
          "SELECT CategoryID, CategoryName FROM dbo.Categories;" +
          "SELECT EmployeeID, LastName FROM dbo.Employees",
          connection);
        connection.Open();
        SqlDataReader reader = command.ExecuteReader();
        while (reader.HasRows)
        {
            Console.WriteLine($"\t{reader.GetName(0)}\t{reader.GetName(1)}");
            while (reader.Read())
            {
                Console.WriteLine($"\t{reader.GetInt32(0)}\t{reader.GetString(1)}");
            }
            reader.NextResult();
        }
    }
}
```

3.2.5 Periodic Forecasting Function

This function uses the ARIMA model to analyze and summarize trends in the attention and consultation volume of legal topics on new media platforms. The model is divided into three components—AR, I, and MA—and analyzed using the variables p, d, and q.The AR component describes the relationship between historical and current values, defining the value of p, which represents the relationship between the current time and the previous p time points. The p parameter is selected using values from the PACF (Partial Autocorrelation Function) plot. The I component represents the order of differencing d, which stabilizes the time series for model training. Unstable time series are transformed into stable series through differencing. The MA component represents the moving average of errors, determining the time difference of the previous q values, which helps in selecting the p parameter.By inputting specific legal topic keywords or consultation categories, the transformed data undergoes differencing to extract stabilized data and determine q and p. Using the ARIMA function to train the model, the system can generate trend charts showing the attention or consultation volume of the topic over a future period, providing foresight on legal hotspots or recommendations for service resource allocation.

4. Comparative Experiment

A comparative experiment was conducted to verify the feasibility of the system proposed in this paper and to compare its advantages with other systems. Naive Bayes and an online analysis system for processing foreign names were selected as control systems for comparison. An I3-3250 CPU application server was configured to compare the response time and operational efficiency of the three analysis systems. Query content was input into the systems, and the resulting response times were recorded. The experimental results are shown in Table 3.As the number of concurrent users increases, the system analysis response time also rises. However, the growth in response time of the system proposed in this paper is relatively small. Under the same number of concurrent users, the analysis response performance of this system is superior, indicating that the system achieves faster analysis response speeds and higher operational efficiency [6].

Table 3. Average Response Time of the Three Systems

Number of Concurrent Users	This System (s)	Control Group A (s)	Control Group B (s)
10	2.254	6.925	5.362
50	3.218	10.521	9.215
100	4.268	22.652	20.248
150	5.165	44.528	41.115
200	5.826	65.254	55.952

5. Conclusion

The natural language information processing system designed in this paper has a wide range of applications. It can extract, analyze, and proofread information from web pages, as well as perform retrieval and forecasting, presenting the processed data results to users. The practical application of the system demonstrates its effectiveness, indicating that the preset functions are valid. During operation, the system can save considerable effort and time at minimal cost, proving its feasibility.

References

[1] Zhang Xiaopeng. Research on Smart Museum Technology Path Based on Unified Data Platform and Large Language Models [J]. Information and Management Research, 2025, 10(03): 1-16.

[2] Sun Hong, Huang Ouyan. Research and Design of Natural Language to Structured Query Statement Algorithm Based on LSTM Fusion [J]. Small and Microcomputer Systems, 2023, 44(01): 63-67.

[3] Suo Wei, Lü Jiaqi, Sun Mengyang, et al. Visual-Linguistic Multimodal Explanation Method Based on Probe Guidance [J]. Journal of Computer Science, 2025, 48(06): 1478-1494.

[4] Wei Ling, Lu Guangyun, Tang Ailong. Personalized Natural Language Sentiment Recognition System Based on Hybrid Neural Networks [J]. Automation and Instrumentation, 2024, 39(09): 26-28+65.

[5] Li Jie, Wang Jizhou, Mao Xi, et al. Semantic Understanding of Natural Language Spatiotemporal Questions Based on Semantic Encoding [J]. Surveying and Mapping Science, 2024, 49(11): 197-206.

[6] Wei Xinling. Design of Automatic Disambiguation Field Recognition System in Natural Language Based on Knowledge Base [J]. Automation Technology and Application, 2023, 42(01): 69-72+151.