Syntactic Analysis of SMOSS Model Combined with Improved LSTM Model: Taking English Writing Teaching as an Example

Ke Yan

xiazi068@126.com

Nanyang Medical College

Article

Keywords: SMOSS model, LSTM model, Syntactic analysis, English writing teaching, Text error detection model, Syntactic structure, Semantic information

Posted Date: September 7th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3233080/v1

Additional Declarations: No competing interests reported.

Abstract

In order to improve learners' syntactic understanding and writing ability, thus effectively improving the quality of English writing teaching, this paper explores the method of combining Sequential Matching on Sliding Window Sequences (SMOSS) model with improved Long Short-Term Memory (LSTM) model in English writing teaching to improve the effect of syntactic analysis. Firstly, this paper analyzes the structure of SMOSS model. Secondly, this paper optimizes the traditional LSTM model by using Connectist Temporal Classification (CTC), and proposes an English text error detection model. Meanwhile, this paper combines the SMOSS model with the optimized LSTM model to form a comprehensive syntactic analysis framework, and designs and implements the structure and code of the framework. Finally, on the one hand, the semantic disambiguation performance of the model is tested by using SemCor data set. On the other hand, taking English writing teaching as an example, the proposed method is further verified by designing a comparative experiment in groups. The results show that: (1) From the experimental data of word sense disambiguation, the accuracy of the SMOSS-LSTM model proposed in this paper is the lowest when the context range is "3 + 3", then it rises in turn at "5 + 5" and "7 + 7", reaches the highest at "7 + 7", and then begins to decrease at "10 + 10"; (2) Compared with the control group, the accuracy of syntactic analysis in the experimental group reached 89.5%, while that in the control group was only 73.2%. (3) In the aspect of English text error detection, the detection accuracy of the proposed model in the experimental group is as high as 94.8%, which is significantly better than the traditional SMOSS-based text error detection method, and its accuracy is only 68.3%. This paper verifies the effectiveness and practicability of applying SMOSS model and improved LSTM model to the syntactic analysis task in English writing teaching, and provides new ideas and methods for the application of syntactic analysis in English teaching.

Introduction

In English writing teaching, syntactic analysis has always been a key step for learners to understand and use grammar rules, which is very important for improving English writing ability [1]. However, many learners are often confused when faced with complex syntactic structures and grammatical rules, which leads to the decline of their writing quality [2]. Therefore, seeking an effective syntactic analysis method is helpful to improve learners' syntactic understanding and writing ability, which is of great significance to improve the quality of English writing teaching.

In order to improve learners' syntactic understanding and writing ability, this paper explores the method of combining Sequential Matching on Sliding Window Sequences (SMOSS) model with improved Long Short-Term Memory (LSTM) model in English writing teaching to enhance the effect of syntactic analysis [3, 4]. SMOSS model is a matching method based on sliding window sequence, which can effectively capture the local features of sentences and provide strong support for syntactic analysis [5]. As a kind of recurrent neural network with memory unit, LSTM model also shows good performance in dealing with natural language tasks [6]. Therefore, the combination of these two models is expected to overcome their respective limitations in syntactic analysis and achieve more accurate and comprehensive syntactic analysis.

The research goal of this paper is to explore the organic integration of SMOSS model and improved LSTM model, and apply it to the syntactic analysis task of English writing teaching [7]. In order to achieve this goal, the structure of SMOSS model is deeply analyzed in this research method part. The traditional LSTM model is optimized by using Connectist Temporal Classification (CTC), and an English text error detection model is proposed. The SMOSS model and the optimized LSTM model are combined to form a comprehensive syntactic analysis framework, and the structure and code of the framework are designed and implemented. In order to verify the effectiveness and practicability of the proposed method, two experiments are used in this paper. SemCor data set is used to test the semantic disambiguation performance of the model. Taking English writing teaching as an example, the proposed method is verified in the actual English writing teaching scene by designing a comparative experiment in groups. The innovation of this paper is that in the process of improving the LSTM model, this paper introduces the connection time series classification model CTC and proposes an English text error detection model. This model helps learners to better understand and apply grammar rules by correcting grammatical errors in learners' texts. The introduction of this model provides new ideas and methods for syntactic analysis in English writing teaching. The contribution of this study lies in exploring the combined application of the two models in the field of syntactic analysis, which provides new ideas and methods for English writing teaching.

Literature review

SMOSS model is a method based on sliding window sequence matching, which can capture the local features of sentences [8]. In recent years, SMOSS model has gradually attracted attention in the field of natural language processing, and has been used in text classification, sentiment analysis and other tasks [9]. In the teaching of English writing, Czischek et al. (2022) began to try to apply SMOSS model to the task of syntactic analysis [10]. The ability to capture local features of SMOSS model brings new possibilities for syntactic analysis. Natraj et al. (2023) found that SMOSS model can effectively analyze the local structure of sentences and help learners better understand the grammatical rules of sentences [11]. For example, in the aspect of syntactic error detection, Ahmed et al. (2022) found that SMOSS model can identify local errors in learners' sentences and provide targeted correction suggestions, thus helping learners improve their writing expression [12]. LSTM is a kind of recurrent neural network (RNN), which is designed with memory units, and can better deal with long-term dependence [13]. In the field of natural language processing, LSTM model performs well in language modelling, machine translation, semantic analysis and other tasks. In the teaching of English writing, Cabra Lopez et al. (2022) began to explore the application of LSTM model to the task of syntactic analysis, in order to improve learners' syntactic understanding and writing ability [14]. The advantage of LSTM model is that it shows better performance when dealing with long text sequences [15]. Ciampelli et al. (2023) found that in English writing teaching, learners often need to understand the structure and grammatical rules of complex sentences, and the memory unit of LSTM model can help learners better capture the long-term dependence in sentences [16].

At present, although the application of SMOSS model and LSTM model in English writing teaching is still in the primary stage, some experimental studies have made some progress. Applying SMOSS model and LSTM model to the task of syntactic analysis in English writing teaching is expected to improve the accuracy and practicability of syntactic analysis. However, there are still some challenges, such as the complexity of the model, the matching of data sets and the handling of error types. Therefore, it is necessary to further study and explore the optimization and improvement of SMOSS model and LSTM model in English writing teaching. Based on the literature review, this paper will continue the exploration of SMOSS model and LSTM model in English writing teaching.

Research on syntactic analysis model

Structural analysis of SMOSS model

The core idea of SMOSS model is to extract local features from sentences by sliding windows, so as to capture local semantic information of sentences [17, 18]. Its structure consists of the following key components: input layer, sliding window, feature extractor, feature matching and output layer [19]. Among them, the core of SMOSS model is matching function, which performs matching operation on sliding window. Assuming that the sliding window size is *w* and the sliding window position is *p*, the matching function $M\left(p\right)$ can be expressed as Eq. (1) [20].

$$M\left(p\right)=f\left(X\left[\,p:p+w\,\right]\right)$$

1

$X[p:p+w]$ represents a word-oriented quantum sequence in a window of length *w* starting from position *p*. The function $f\left(\cdot\right)$ is a nonlinear mapping function, which is used to map the words in the window to the quantum sequence to the matching representation space.

In order to capture the relationship between matching representations in different positions, SMOSS model introduces matching attention mechanism. Assuming that the attention weight between matching representations is *A*, when the *i*-th matching representation $M\left(p_i\right)$ is considered, its weighted representation is Eq. (2) [21].

$$M_{\text{weighted }}\left(p_i\right)=\sum_{j=1}^{N}A_{ij}\cdot M\left(p_j\right)$$

2

A_{ij} represents the attention weight between the *i*-th matching representation and the *j*-th matching representation.

Input the weighted representation of the matching representation into a fully connected layer to obtain the overall SMOSS representation, as shown in Eq. (3) [22].

$$SMOSS_{\text{output}} = \operatorname{ReLU}\left(W \cdot M_{\text{weighted}} + b\right)$$

3

W and *b* are the parameters of the fully connected layer, and ReLU represents the activation function.

The specific structure of SMOSS is shown in Fig. 1.

Based on the content in Fig. 1, the input layer of SMOSS model is responsible for receiving the text data to be processed. In the task of parsing, the input layer transforms English sentences into word sequences and embeds them, that is, each word is mapped into a dense vector representation. Such an embedded representation can better represent the semantic relationship between words. Sliding window is one of the core components of SMOSS model. It cuts the input word sequence into several sub-sequences with fixed length, and then performs feature extraction and matching on each sub-sequence [23]. The size of sliding window is an important super parameter, which determines the range of local features.

Optimization of improved LSTM model based on CTC

CTC is an end-to-end sequence learning method, which is widely used in sequence labelling tasks. The basic structure of CTC is shown in Fig. 2.

In this paper, CTC is applied to the task of syntactic analysis, and the LSTM model is optimized by CTC loss function to improve the syntactic analysis effect in English writing teaching [24].

In the sequence labelling task, given the input sequence $X = (x_1, x_2, \ldots, x_T)$, it is necessary to predict the output sequence $Y = (y_1, y_2, \ldots, y_T)$. *T* represents the length of the input sequence. *Y* represents the length of the output sequence. However, because the sequence length of different samples may be different, the traditional labelling data usually need to be strictly aligned, that is, the input sequence and the output sequence are required to have the same length [25]. This will be a challenge for the task of syntactic analysis, because learners' sentence structures are diverse, resulting in different sentences with different lengths.

The key of CTC principle is to solve the sequence alignment problem by defining blank symbol Ø and repeated symbol *r*. Assuming that the input sequence is "hello", the possible output sequences are "healo", "helo" or "hello". The goal of CTC model is to find all possible alignments and calculate the probability of each alignment. In order to achieve this, CTC introduces blank symbols in the output sequence *Y*, indicating that there may be silent blank parts, such as "heØaØllo". The repetition symbol *r* is used to represent the same characters in succession, such as "heØarØllo". By introducing blank symbols and repeated symbols, CTC can find all possible alignment paths [26].

When calculating the CTC loss function, it is necessary to accumulate all possible alignment paths to get the difference between the output sequence *Y* and the real output sequence [27]. Minimizing the loss function will make the model better adapt to the mapping relationship between input sequence and output sequence, thus optimizing the whole LSTM model.

Given the input sequence $X = (x_1, x_2, \ldots, x_T)$ and the output sequence $Y = (y_1, y_2, \ldots, y_T)$, the CTC loss function is defined as Eq. (4).

$$L_{CTC} = -\log \sum_{\pi \in B^{-1}(Y)} P(\pi \mid X)$$

4

$B^{-1}\,(Y)$ represents the set of all possible alignment paths of the output sequence *Y*. $P(\pi \mid X)$ represents the conditional probability of output sequence π given input sequence *X*. The goal of the loss function is to minimize L_{CTC} to optimize the model parameters and make the difference between the predicted output sequence and the real output sequence as small as possible.

By introducing blank symbols and repeated symbols, CTC's flexible sequence alignment mechanism enables the improved LSTM model to learn without complete alignment labels and better adapt to the diverse sentence structures in learners' writing expressions. This provides new ideas and methods for the application of syntactic analysis in English writing teaching, and lays the foundation for the innovation of this study.

In order to apply CTC method to improve the optimization of LSTM model, the CTC loss function is connected with the output layer of LSTM. In the traditional LSTM model, the output layer usually maps the hidden state of LSTM to the classification label space through the fully connected layer. However, in this paper, the output sequence of LSTM model is directly used as the input sequence of CTC without introducing additional full connection layer.

In LSTM model, the hidden state sequence $H = (h_1, h_2, \ldots, h_T)$ is obtained by recursive calculation. h_T represents the hidden state at time *T*, and the output sequence $O = (o_1, o_2, \ldots, o_T)$ of LSTM model. The traditional LSTM model will use the fully connected layer to map the hidden state to the target label space, and then carry out the classification task. However, in this paper, the output sequence of LSTM model is directly used as the input sequence of CTC without introducing additional full connection layers.

The combination mode of the improved LSTM model based on CTC is shown in Table 1 [28]:

Table 1
Combination steps of improved LSTM model based on CTC

Step number	Specific content
1	Input the input sequence *X* into the LSTM model, and get the hidden state sequence $H = (h_1, h_2, \ldots, h_T)$ through recursive calculation.
2	The output sequence $O = (o_1, o_2, \ldots, o_T)$ of the LSTM model is associated with the hidden state *H*, and the output sequence *O* is taken as the input sequence of CTC.
3	CTC model uses blank symbol Ø and repeated symbol *r* to find all possible alignment paths and learn the corresponding relationship between output sequences and real labels.
4	Minimize the CTC loss function to optimize the parameters of the whole model, that is, $min_{\theta}L_{CTC}\ (\theta)$, where θ represents the parameters of the model.

The goal of the training process of the improved LSTM model based on CTC is to optimize the model parameters by minimizing the CTC loss function to achieve better syntactic analysis effect. In the training process, the whole model is jointly trained by using back propagation algorithm combined with CTC loss function. Let $P(Y \mid X)$ be the probability between the output sequence o of CTC model and the real output sequence *Y*, and the CTC loss function is defined as $L_{CTC} = -\log P(Y \mid X)$. The parameters of the whole model are optimized by minimizing CTC loss.

In the training process, the Stochastic Gradient Descent (SGD) algorithm is used for optimization, and the learning rate adjustment strategy is used to speed up the convergence of the model. In addition, in order to prevent over-fitting, techniques such as Dropout and L2 regularization are introduced. Specifically, for each sample (X, Y), the gradient $\nabla_{\theta}L_{CTC}(X, Y; \theta)$ of its CTC loss function with respect to parameter θ is calculated. Then, the learning rate α is used to update the parameters, and the updating formula is shown in Eq. (5) [29].

$$\theta \leftarrow \theta - \alpha \cdot \nabla_{\theta}L_{CTC}(X, Y; \theta)$$

5

α is the learning rate, which is used to control the pace of parameter update.

In order to prevent the model from over-fitting, Dropout technology is introduced, which can randomly set the output of some neurons to zero, thus reducing the dependence between neurons and enhancing the generalization ability of the model [30]. Meanwhile, L2 regularization is also adopted, and the L2 norm of parameters is introduced into the loss function to suppress the situation that the parameters are too large, thus further preventing the model from over-fitting the training data [31]. Dropout technology reduces the dependence between neurons and enhances the generalization ability of the model by randomly setting the output of some neurons to zero. Let h_i represent the output of the ith neuron, and in the training process, a retention probability p is used to control the retention and discarding of neurons, that is, Eq. (6):

$$h_i = \begin{cases} h_i \cdot r_i, \text{with probability } p \\ 0, \text{with probability } 1 - p \end{cases}$$

6

r_i is a random number that obeys the uniform distribution U(0,1). During the test, the output of all neurons is kept instead of using Dropout.

R_i is a random number that obeys the uniform distribution U (0,1). During the test, the output of all neurons is kept instead of using Dropout.

L2 regularization suppresses the condition that the parameters are too large by introducing L2 norm into the loss function, thus further preventing the model from over-fitting the training data. The calculation process is shown in Eq. (7).

$$R\left(\theta\right) = \sum_{i=1}^{N} \theta_i^2$$

7

θ is the model parameter. N is the total number of model parameters. The L2 regularization term is added to the original CTC loss function to obtain the regularized loss function, as shown in Eq. (8).

$$L_{\text{reg}} = L_{CTC} + \lambda \cdot R\left(\theta\right)$$

8

λ is a regularization parameter used to control the regularization intensity.

Design and implementation of comprehensive syntactic analysis framework

In the framework of comprehensive syntactic analysis, the SMOSS model is integrated with the optimized LSTM model. Specifically, firstly, the SMOSS model is used to capture the local features of the input text and obtain the coded representation of the sliding window subsequence. Then, these encoded representations are input into the optimized LSTM model, and the encoded representation of the whole text is obtained by LSTM encoder.

In order to make better use of the context information between SMOSS model and LSTM model, a matching layer is introduced into the framework. The matching layer adopts attention mechanism, and realizes the interaction between local features and global features by calculating the similarity between each sliding window subsequence and other subsequence. Specifically, assuming that there are *N* sliding window subsequence in the input text, the output of the matching layer can be expressed as Eq. (9).

$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

9

Q,K,V respectively represent the local features of SMOSS model output, the global features of LSTM model output and the features that need to be fused, and d_k is the feature dimension.

In order to better cope with the task of error detection in English writing teaching, the LSTM model is extended and a sequence-to-sequence (seq2seq) model is constructed. The model includes an encoder and a decoder, which are used to encode the input text and generate the corrected text respectively.

In the encoder part, the improved LSTM model is used to encode the input text, and the coding representation *C* is obtained. In the decoder part, another LSTM model is used to generate the corrected text step by step. Assuming that the target text is *Y*, the goal of the decoder is to maximize the conditional probability of generating the target text, that is, Eq. (10).

$$P(Y\mid X)=\prod_{t=1}^{T}P\left(y_t\mid y_1,y_2,\ldots,y_{t-1},C\right)$$

10

T represents the target text length.

In the implementation of the comprehensive parsing framework, the SMOSS model and the improved LSTM model are realized by using Python programming language and TensorFlow, a deep learning framework. Batch training and optimizer are used to train the model, and cross entropy loss function is used to measure the difference between the prediction results of the model and the real label.

Analysis of the evaluation results of syntactic analysis model effect

Data set introduction

Firstly, the SemCor data set is selected as a main evaluation data set in this study. SemCor is a semantic contextual tagging corpus created by Princeton University, which contains a large number of semantic tagging of English sentences and words. This data set is widely used in natural language processing tasks such as word sense disambiguation and syntactic analysis. This article uses 1,000 data in the SemCor dataset.

In addition to SemCor data set, in order to be closer to the actual scene of English writing teaching, an English writing teaching data set is also built. This data set contains sentences and paragraphs written by learners in English writing teaching, including examples with different grammatical difficulty and writing level. The sentences in the data set include all kinds of grammatical errors, such as inconsistent subject and predicate, tense errors, improper use of articles and so on. A total of 500 self-built datasets are collected.

Analysis of model performance evaluation results

The test set and test set of SemCor data set are divided into 8:2, and the comprehensive syntactic analysis framework model constructed in this paper is trained and evaluated. Taking the traditional SMOSS-based text error detection method as the control group and the comprehensive syntactic analysis framework model in this paper as the experimental group, the evaluation results of the model performance are compared, and the results are shown in Table 2 for three experiments.

Table 2
Model performance evaluation results

Number of experiments	Group	Accuracy	Recall rate	F1 score
1	Experimental group	0.890	0.870	0.882
	Control group	0.730	0.711	0.721
2	Experimental group	0.891	0.870	0.882
	Control group	0.731	0.712	0.723
3	Experimental group	0.890	0.873	0.884
	Control group	0.732	0.711	0.720

The visualization result is shown in Fig. 3.

Figure 3 shows that the accuracy of the experimental group is between 0.890 and 0.891, the recall rate is between 0.870 and 0.893, and the F1 score is between 0.882 and 0.884. The accuracy of the control group is 0.7030 at the minimum and 0.732 at the maximum, and the recall rate and F1 score are also around 0.710. According to these data, it can be concluded that the experimental group performs better in syntactic analysis tasks than the control group.

Analysis of experimental results of word sense disambiguation

SemCor data set is used to test the semantic disambiguation performance of the model to evaluate the effect of the improved SMOSS-LSTM model in disambiguation. Several ambiguous words are selected as test samples, including "bank", "plant" and other common ambiguous words. By embedding these words into sentences, test data with different contexts are constructed and input into the improved SMOSS-

LSTM model for disambiguation. The comparison of disambiguation accuracy between the experimental group and the control group is shown in Table 3.

Table 3
Experimental results of word sense disambiguation

Number of experiments	Group	3 + 3	5 + 5	7 + 7	10 + 10
1	Experimental group	0.821	0.870	0.890	0.831
	Control group	0.760	0.811	0.852	0.780
2	Experimental group	0.792	0.850	0.881	0.812
	Control group	0.753	0.791	0.841	0.763
3	Experimental group	0.810	0.862	0.901	0.820
	Control group	0.741	0.780	0.851	0.772

The diagram is drawn according to Table 3, as shown in Fig. 4.

According to Fig. 4, by comparing the results of the experimental group and the control group, it can be found that the improved SMOSS-LSTM model in the experimental group is superior to the benchmark model in the control group in disambiguation accuracy. The improved SMOSS-LSTM model makes better use of contextual information, especially in a larger context, and can understand sentence context more comprehensively, thus eliminating ambiguity more accurately. However, the benchmark model in the control group is not as sensitive to the use of context information as the improved SMOSS-LSTM model, resulting in low accuracy.

Validation of syntactic analysis effect and analysis of comparative experimental results

In the verification and comparison of syntactic analysis effect, the control group is set as the traditional dependency parser. The experimental group is the improved SMOSS-LSTM model. The experiments are still carried out with sentences in SemCor data set, and different sample sizes (50, 100, 150) are selected to investigate the generalization ability and effect of the model under different data sizes. The results are shown in Table 4.

Table 4
Validation of syntactic analysis effect and comparison of experimental results

Number of experiments	Group	50	100	150
1	Comprehensive analysis framework	0.850	0.895	0.821
	Traditional dependency parser	0.711	0.741	0.721
2	Comprehensive analysis framework	0.860	0.882	0.820
	Traditional dependency parser	0.732	0.711	0.722
3	Comprehensive analysis framework	0.870	0.891	0.840
	Traditional dependency parser	0.732	0.720	0.712

The diagram is drawn according to Table 4, as shown in Fig. 5.

According to the data in Fig. 5, compared with the control group, the accuracy of syntactic analysis in the experimental group reaches 89.5%, while that in the control group is only 73.2%. With the increase of sample size, the accuracy of the experimental group and the control group fluctuated, but the overall trend showed good stability. Because of the combination of SMOSS model and improved LSTM model, a comprehensive syntactic analysis framework is formed, which makes full use of their advantages and improves the effect of syntactic analysis.

Analysis of experimental results of English text error detection

In this paper, the traditional SMOSS-based text error detection method and the improved SMOSS-LSTM are used to carry out the comparative experiment of English text error detection, and the self-built data set is used to carry out the experiment for three times. The results are shown in Table 5.

Table 5
Experimental results of English text error detection

Number of experiments	Group	50	100	150
1	Comprehensive analysis framework	0.930	0.942	0.948
	Text error detection based on SMOSS	0.661	0.651	0.683
2	Comprehensive analysis framework	0.922	0.930	0.944
	Text error detection based on SMOSS	0.633	0.651	0.667
3	Comprehensive analysis framework	0.911	0.931	0.947
	Text error detection based on SMOSS	0.600	0.632	0.681

The diagram is drawn according to Table 5, as shown in Fig. 6.

In Fig. 6, in the aspect of English text error detection, the detection accuracy of the proposed model in the experimental group is as high as 94.8%, which is significantly better than the traditional SMOSS-based text error detection method, and its accuracy is only 68.3%. In the experiment of English text error detection, the proposed SMOSS-LSTM model has a high detection accuracy in the experimental group, which is significantly better than the traditional SMOSS-based text error detection method. It also shows that with the increase of sample size, the accuracy of the experimental group tends to be stable, while the accuracy of the control group changes little. This further verifies the stability and effectiveness of SMOSS-LSTM model in dealing with complex sentences and long texts.

Discussion

From the experimental results, whether the sample size is 50, 100 or 150, the accuracy of the experimental group (using SMOSS-LSTM model) remains at a high level, and the highest is 94.8%. This shows that the proposed SMOSS-LSTM model has obvious advantages in English text error detection, and verifies its application potential in English writing teaching. In contrast, the control group of the traditional text error detection method based on SMOSS has a low accuracy, and the highest accuracy is only 68.3%. This further verifies the significant improvement of the proposed model compared with the traditional method. From the experimental results, it is also observed that the accuracy of the experimental group changes little under different sample sizes, which shows that the proposed SMOSS-LSTM model is stable when dealing with data of different sizes. This shows that the model has good adaptability to English texts with different lengths and complexity, and can be applied to a variety of practical scenarios, including English writing teaching for learners, natural language processing and other fields. The experimental results also show that the accuracy of the experimental group tends to be stable with the increase of sample size. This phenomenon may be related to the characteristics of LSTM model, which has advantages in dealing with long-term dependence.

Conclusions

In this paper, the method of combining SMOSS model with improved LSTM model is explored and applied to the syntactic analysis task in English writing teaching. The experimental results verify the effectiveness and advantages of the proposed SMOSS-LSTM model in English text error detection. By comparing the results of the experimental group and the control group, it is observed that the accuracy of the experimental group is significantly higher than that of the control group. The experimental results show that with the increase of sample size, the accuracy of the experimental group tends to be stable, which shows that the proposed SMOSS-LSTM model is stable when dealing with data of different sizes, and has the advantage of adapting to different text lengths and complexity. However, it is also noted that there are some shortcomings in this paper. First, the sample size of experimental data is relatively small, so people can consider expanding the sample size in the future to enhance the reliability of experimental results. Secondly, although the experimental data set covers common types of grammatical errors, there may be other types of errors that need further consideration. In addition, this paper focuses on English text error detection, and the model can be applied to other natural language processing tasks for a more comprehensive evaluation in the future.

Declarations

Data availability

All data generated or analyzed during this study are included in this published article [and its supplementary information files].

References

1. Zhao P, Lu W, Wang S, Peng X, Jian P, Wu H, et al. Multi-granularity interaction model based on pinyins and radicals for Chinese semantic matching. World Wide Web. 2022; 25(4): 1703–1723.
2. Wang H, Li J, Wu H, Hovy E, Sun Y. Pre-trained language models and their applications. Engineering. 2022; 15: 112.
3. Chiao C S, Lin D T. ELICE: Embedding Language through Informative Contrastive-Encoder. Expert Systems with Applications. 2023; 229: 120523.
4. Sadeghi F, Bidgoly A J, Amirkhani H. Fake news detection on social media using a natural language inference approach. Multimedia Tools and Applications. 2022; 81(23): 33801–33821.
5. Wu Z. Neural Fuzzy Logic Reasoning for Natural Language Inference. 2022; 21: 68.
6. Kobayashi G, Tanaka K F, Takata N. Pupil dynamics-derived sleep stage classification of a head-fixed mouse using a recurrent neural network. The Keio Journal of Medicine. 2023; 72(2): 44–59.
7. Zhao P, Lu W, Wang S, Peng X, Jian P, Wu H, et al. Multi-granularity interaction model based on pinyins and radicals for Chinese semantic matching. World Wide Web. 2022; 25(4): 1703–1723.
8. Wang X, Alonso-Mora J, Wang M. Probabilistic risk metric for highway driving leveraging multi-modal trajectory predictions. IEEE Transactions on Intelligent Transportation Systems. 2022; 23(10): 19399–19412.
9. Cruz M F, Ono N, Huang M, Altaf-Ul-Amin M, Kanaya S, Cavalcante C A M T. Kinematics approach with neural networks for early detection of sepsis (KANNEDS). BMC Medical Informatics and Decision Making. 2021; 21(1): 1–11.
10. Czischek S, Moss M S, Radzihovsky M, Merali E, Melko R G. Data-enhanced variational Monte Carlo simulations for Rydberg atom arrays. Physical Review B. 2022; 105(20): 205108.
11. Natraj S, Kojovic N, Maillart T, Schaer M. Video-Audio Neural Network Ensemble For Comprehensive Screening Of Autism Spectrum Disorder in Young Children. medRxiv. 2023; 2023.06. 28.23291938.
12. Ahmed M I B, Alotaibi S, Dash S, Nabil M, AlTurki A O. A review on machine learning approaches in identification of pediatric epilepsy. SN Computer Science. 2022; 3(6): 437.

13. Harvey B J, Olah V J, Aiani L M, Rosenberg L I, Pedersen N P. Classifier for the Rapid Simultaneous Determination of Sleep-Wake States and Seizures in Mice. bioRxiv. 2023; 2023.04. 07.536063.

14. Cabra Lopez J L, Parra C, Gomez L, Trujillo L. Sex recognition through ECG signals aiming toward smartphone authentication. Applied Sciences. 2022; 12(13): 6573.

15. Kojovic N, Natraj S, Mohanty S P, Maillart T, Schaer M. Using 2D video-based pose estimation for automated prediction of autism spectrum disorders in young children. Scientific Reports. 2021; 11(1): 15069.

16. Ciampelli S, de Boer J N, Voppel A E, Corona Hernandez H, Brederoo S G, van Dellen E, et al. Syntactic network analysis in schizophrenia-spectrum disorders. Schizophrenia Bulletin. 2023; 49(Supplement_2): S172-S182.

17. Das A, Ahmed M M. Structural equation modeling approach for investigating drivers' risky behavior in clear and adverse weather using SHRP2 naturalistic driving data. Journal of Transportation Safety & Security. 2022; 18: 1–32.

18. Yuan X, Sunyer-Pons N, Terrado A, León J L, Hadziioannou G, Cloutet E, et al. 3D-Printed Organic Conjugated Trimer for Visible-Light-Driven Photocatalytic Applications. ChemSusChem. 2023; 2023: e202202228.

19. Phillips B N, Iwanaga K, Rumrill S, Reyes A, Wu J R, Fleming A R, et al. Development and validation of the social motivation scale in people with disabilities. Rehabilitation Psychology. 2021; 66(4): 589.

20. Huang B, Zhang J, Ju J, Guo R, Fujita H, Liu J. CRF-GCN: An effective syntactic dependency model for aspect-level sentiment analysis. Knowledge-Based Systems, 2023; 260: 110125.

21. Xiang Y, Kong Y, Feng W, Ye X. Liu Z. A ratiometric photoelectrochemical microsensor based on a small-molecule organic semiconductor for reliable in vivo analysis. Chemical Science. 2021; 12(39): 12977–12984.

22. Ward B J, Andriessen N, Tembo J M, Kabika J, Grau M, Scheidegger A, et al. Predictive models using "cheap and easy" field measurements: Can they fill a gap in planning, monitoring, and implementing fecal sludge management solutions? Water Research. 2021; 196: 116997.

23. Kizawi A, Borsos A. A Literature review on the conflict analysis of vehicle-pedestrian interactions. Acta Technica Jaurinensis. 2021; 14(4): 599–611.

24. Zhang H, Wang Y, Yang Y. Salient-Aware Multiple Instance Learning Optimized Network for Weakly Supervised Object Detection. Available at SSRN. 2021; 12: 4292749.

25. Osella S, Wang M, Menna E, Gatti T. Lighting-up nanocarbons through hybridization: Optoelectronic properties and perspectives. Optical Materials: X. 2021; 12: 100100.

26. Zha W, Liu Y, Wan Y, Luo R, Li D, Yang S, et al. Forecasting monthly gas field production based on the CNN-LSTM model. Energy. 2022; 2022: 124889.

27. Shahid F, Zameer A, Muneeb M. A novel genetic LSTM model for wind power forecast. Energy. 2021; 223: 120069.

28. Behera R K, Jena M, Rath S K, Misra S. Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. Information Processing & Management. 2021; 58(1): 102435.

29. Tiwari S, Jain A, Sapra V, Koundal D, Alenezi F, Polat K, et al. A smart decision support system to diagnose arrhythmia using ensembled ConvNet and ConvNet-LSTM model. Expert Systems with Applications. 2023; 213: 118933.

30. Fan D, Sun H, Yao J, Zhang K, Yan X, Sun Z. Well production forecasting based on ARIMA-LSTM model considering manual operations. Energy. 2021; 220: 119708.

31. Tasdelen A, Sen B. A hybrid CNN-LSTM model for pre-miRNA classification. Scientific reports. 2021; 11(1): 14125.

Figures

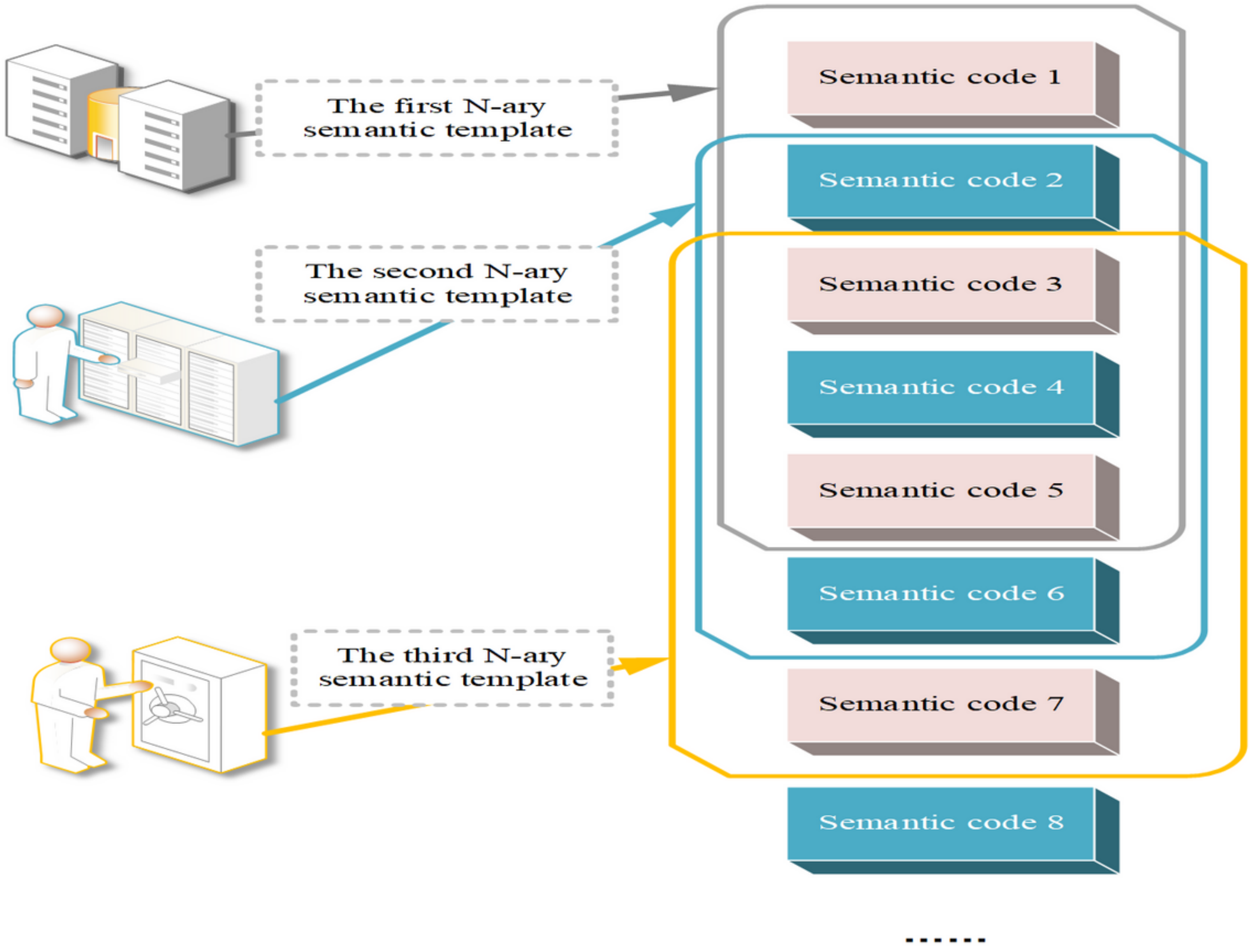

Figure 1

Basic structure of SMOSS model

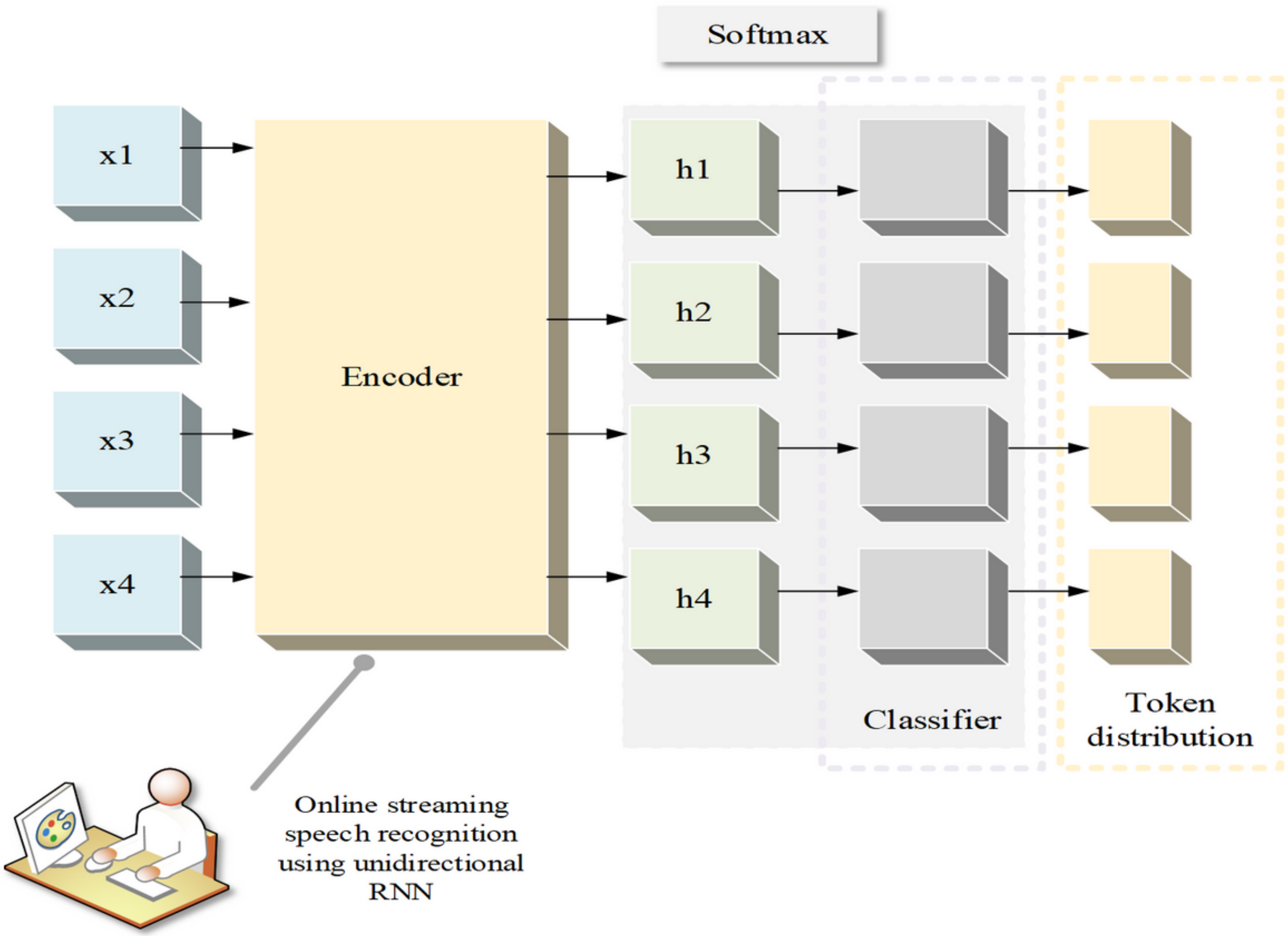
Softmax
x1
x2
x3
x4
Encoder
h1
h2
h3
h4
Classifier
Token distribution
Online streaming speech recognition using unidirectional RNN

Figure 2

Basic structure of CTC model

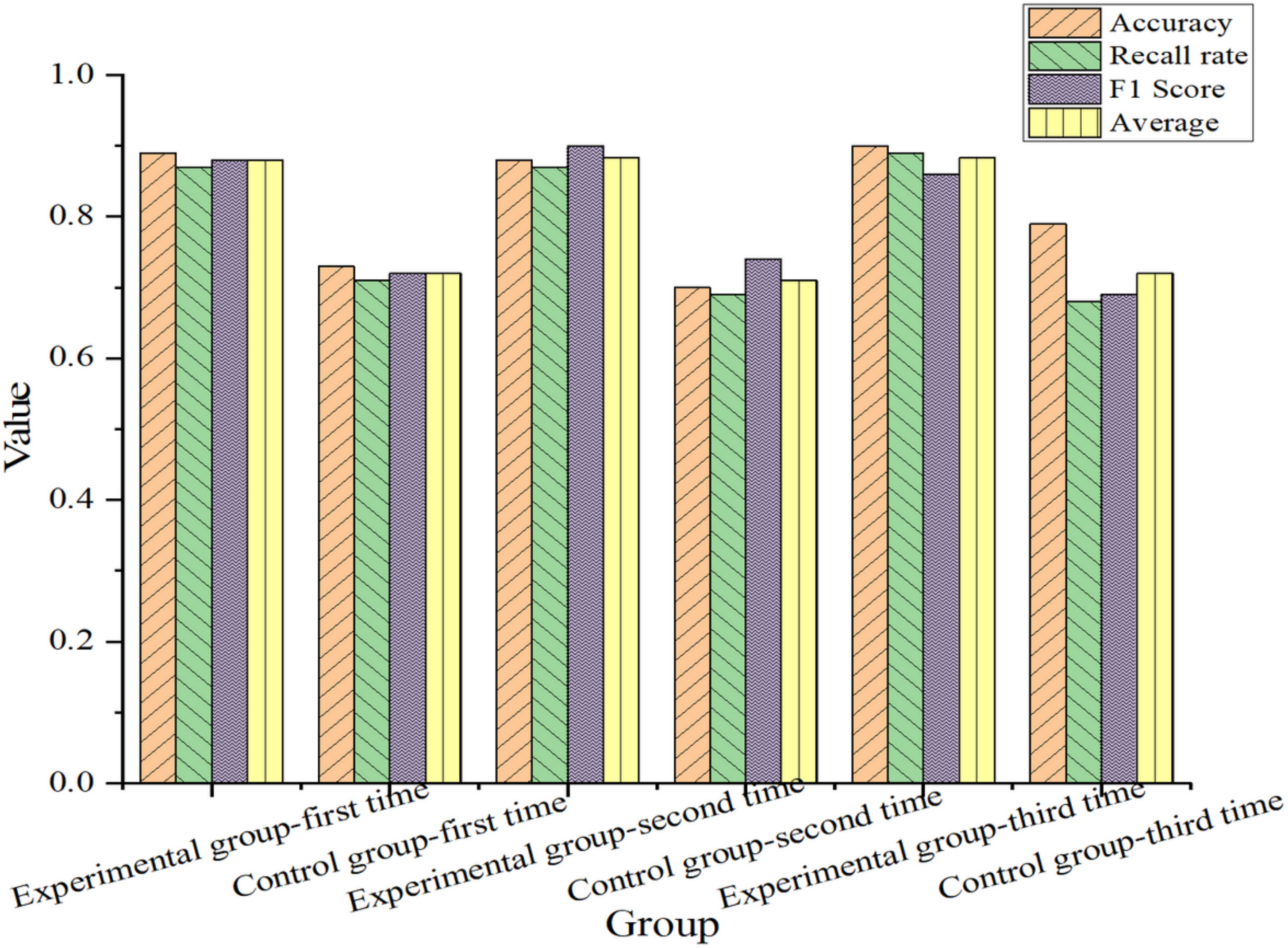

Figure 3

Comparison chart of model evaluation results

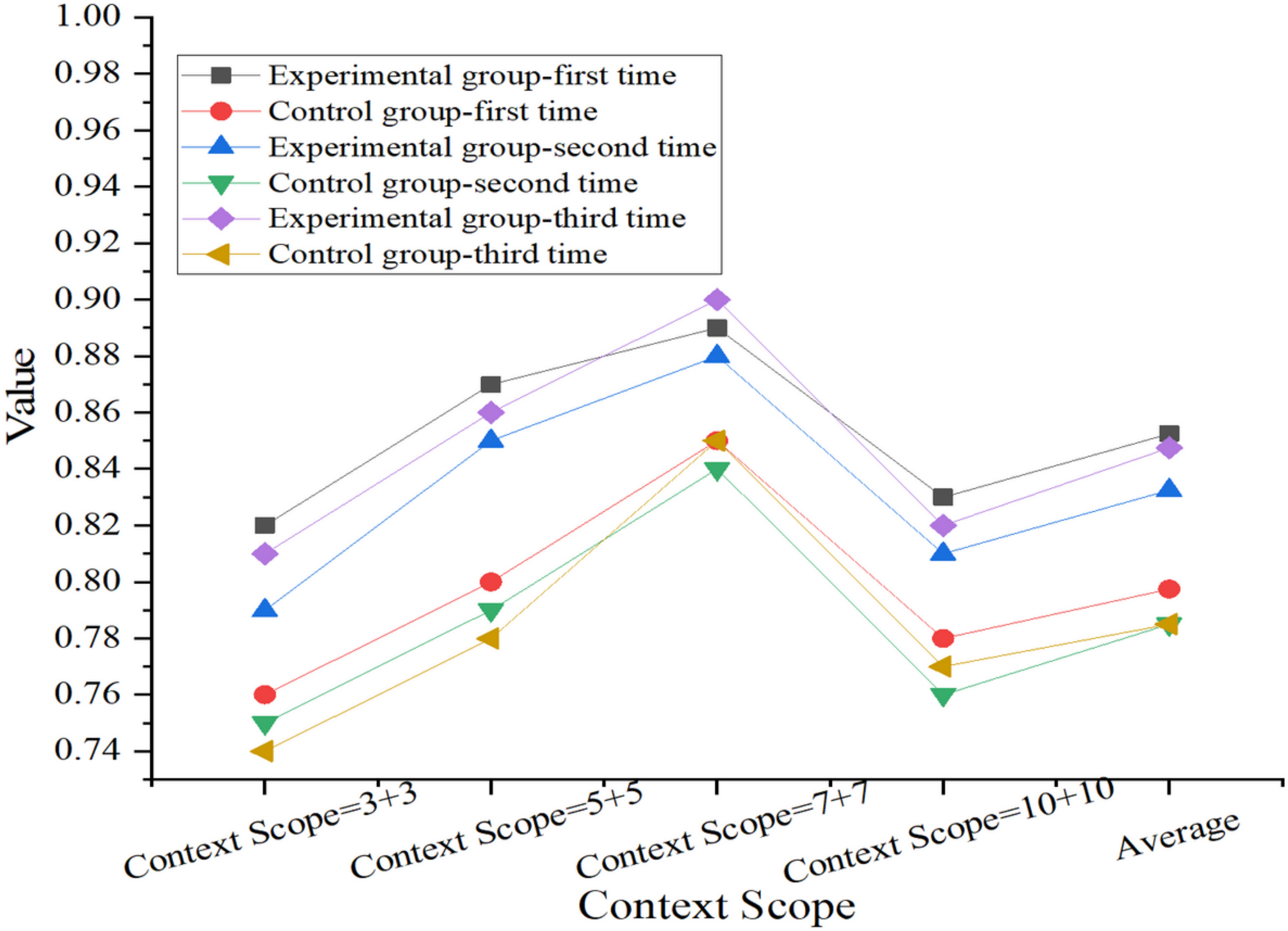

Figure 4

Comparison chart of word sense disambiguation results

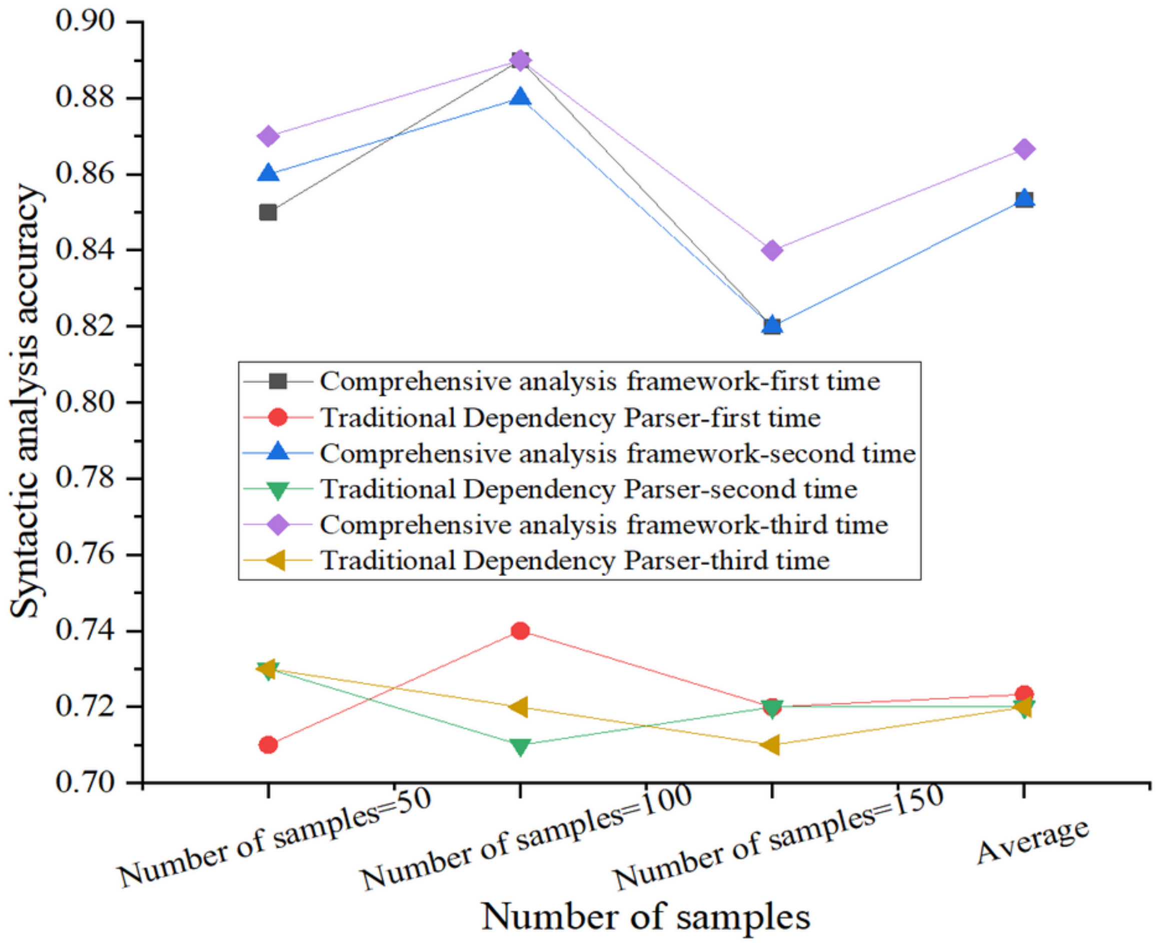

Figure 5

Comparison chart of syntactic analysis effect and result

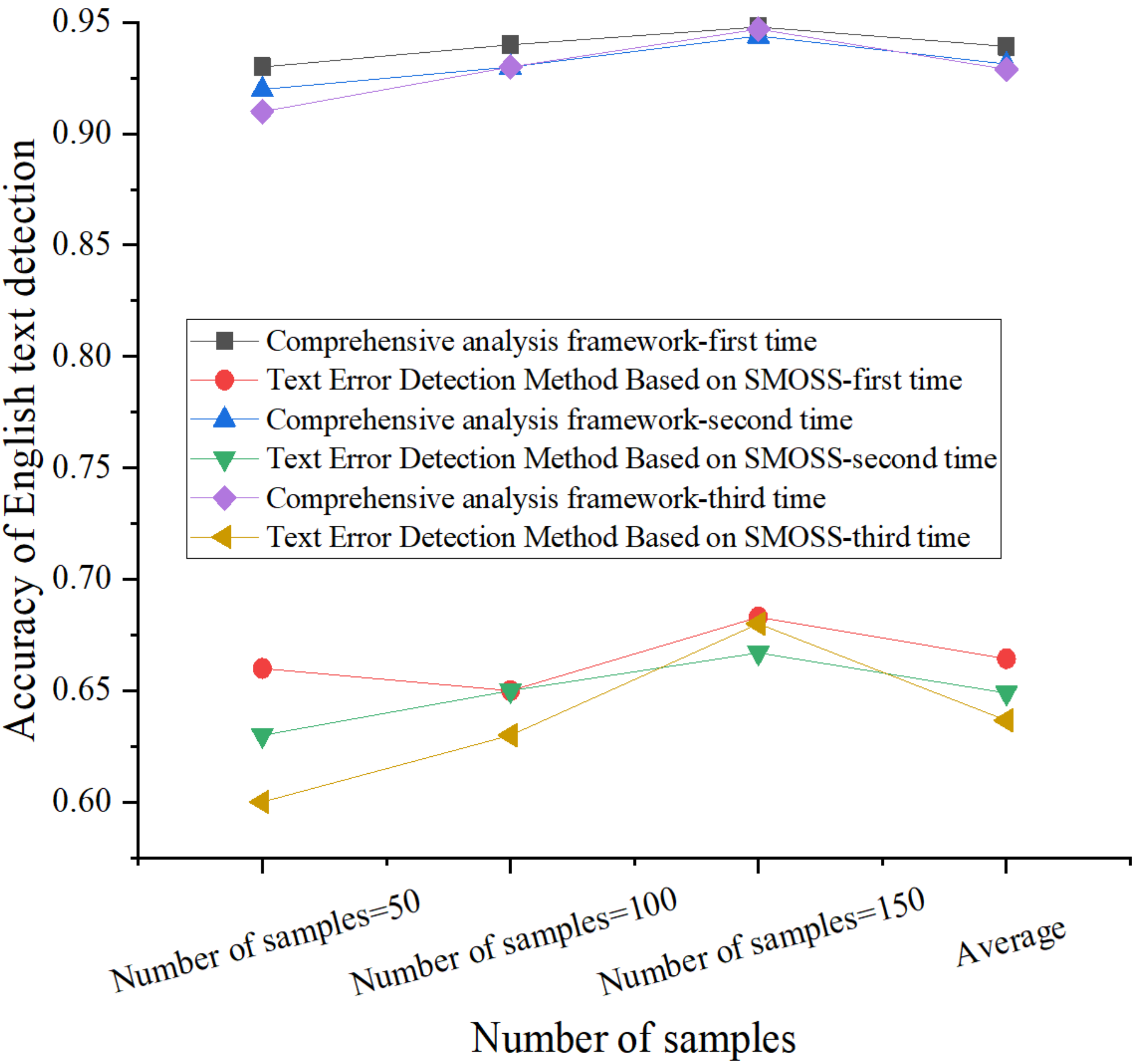

Figure 6

Comparison chart of English text error detection results

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Data1.zip