Intelligent Computing Technology and Automation
Z. Hou (Ed.)

doi:10.3233/ATDE231210

Syntax Error Detection in English Text Images Based on Sparse Representation

Xixi CHEN[a,1]
[a] *School of Foreign Languages, Leshan Normal University, Leshan, 614000, China*

Abstract. In order to solve the problem of poor grammar error detection effect in English text images, a grammar error detection method in English text images based on sparse representation is proposed. The characteristic data of English text images are collected, and the English text image evaluation system is constructed. Combined with the principle design of sparse representation and the grammar error recognition algorithm, it is verified by experiments, the grammar error detection method in English text images based on sparse representation has high practicability and fully meets the research requirements.

Keywords. Sparse representation; English text; Text image; Syntax error

1. Introduction

The detection of grammatical errors in English text images is time-consuming and laborious. Correcting each student's composition one by one makes the task more arduous [1]. At present, to correct the errors in English composition and give correction suggestions can reduce the burden of teachers and directly promote students to participate in writing practice, so as to improve students' English writing level [2]. The complexity of English grammar is relatively high, which also leads to the diversification of grammatical error types. Most of the common grammatical errors are caused by these reasons. Most English beginners often have confused use when using sound like words or shape like words due to their weak basic knowledge. Even the misuse of word formation and the deviation of semantic understanding will lead to grammatical errors [3]. At present, improper sentence collocation is the main type of semantic level errors, including the collocation between subject and predicate, the collocation between verb and object, the collocation between adjective and adverb, and the collocation between quantifier and noun. Based on the level of syntactic structure analysis, although some sentences can make sense, there are ambiguities at the semantic level. The traditional grammar error detection based on word granularity and word granularity has made some progress, and the error correction based on grammar has also made good progress [4]. However, the research on semantics still puzzles researchers until the application of sparse representation technology breaks this deadlock. Before the advent of sparse representation model, the realization of syntax error correction depends on a relatively simple error correction model, and the capture of high-level semantic features is often impossible. With the advent of the era of big data, the training of sparse representation model by large-scale corpus has greatly improved the effect of syntax error correction, the capture of semantic information has become more accurate, and great progress has been made in semantic error correction.

[1] Corresponding Author: Xixi CHEN; School of Foreign Languages, Leshan Normal University, Leshan, 614000, China; chen_xixi19832008@126.com

2. Syntax Error Detection in English Text Images

2.1 Syntax Recognition in English Text Images

In text analysis, part of speech is one of the most commonly used features in English text images. English part of speech refers to the grammatical test results according to the grammatical characteristics and lexical meaning of words, such as nouns, verbs, adjectives and so on. The process of dividing the correct part of speech for the words in the sentence by algorithm is called part of speech tagging [5]. The difficulty of part of speech tagging for different languages is also different. English words often contain multiple parts of speech, which can only be confirmed in the context. However, the part of speech of English words is relatively single, and most English words have only one part of speech. The flow chart of using sparse representation to check syntax errors is shown in the figure 1 below.

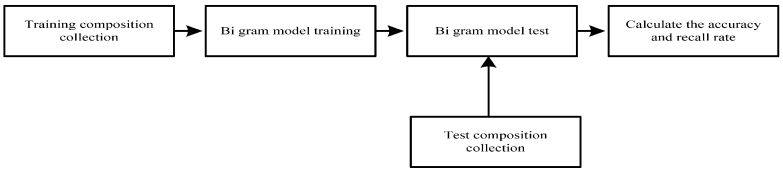

Figure 1. Real word error checking process of English text image

At present, there are many open source part of speech tagging tools, such as thulac developed by natural language processing and social humanities Computing Laboratory Based on sparse representation [6], This study uses the LTP language technology platform to label the part of speech of the text corpus. LTP uses 863 words, and the meanings of their parts of speech are shown in the table 1.

Table 1. Part of speech meaning of English text

Label	Meaning	Give an example	Label	Meaning	Give an example
a	adjective	Intelligence	pw	Organization name	China Telecom
b	Other rhetorical names	Chinese style	pi	Positional NOUN	suburb
c	conjunction	because	o	place name	Shanghai
d	adverb	very	p	Tense NOUN	today
e	Interjection	ah	q	Other proper nouns	the United Nations
f	morpheme	Ci	r	an onomatopoeia	Ding Dong
g	prefix	false	s	preposition	stay
h	idiom	the wind is mild and the sun is bright	t	quantity	individual
i	abbreviation	Olympic Games	u	pronoun	they
j	suffix	rate	v	auxiliary word	land
k	number	five	w	verb	run
l	General NOUN	desk	x	punctuation	?
m	Location NOUN	right	y	Loanwords	WTO
n	name	Du Fu	z	Non constituent word	Soar

Word posteriori probability is a commonly used confidence feature in statistical machine translation error detection. The confidence of the target word can be expressed

by a posteriori probability, that is, $\left(e_{n,i}, e\right)$ represents an SMT. Given the source language input sentence p is the corresponding sparse representation list output δ, where $\left(f_1^J, e_{n,1}^{n,l_n}\right)$ stands for the nth translation hypothesis in the N-best list, the translation probability of each translation hypothesis is recorded as $\left(e_{n,i}, e^{'}\right)$, and the WP used for error detection is the posterior probability p(E1//1) of each target word in the 1est list. Three typical methods for calculating WPP are described in detail below. Given the source language input F, the a posteriori probability of the target word at position I in the optimal translation hypothesis E is the sum of the probabilities of other translation hypotheses in the N-best list at the corresponding fixed position D, as shown in the formula:

$$p_i\left(e \mid f_1^J\right) = \frac{F\sum_{n=1}^{N}\delta\left(e_{n,i}, e\right)}{\sum_{e}\sum_{n=1}^{N}\delta\left(e_{n,i}, e^{'}\right)\cdot p\left(f_1^J, e_{n,1}^{n,l_n}\right)} - \text{E D} \tag{1}$$

In the sparse representation list, the sentence length of translation assumptions changes dynamically within a certain range, so that due to the influence of different translation assumptions, the target word E in the same position I may be different, that is, corresponding to different source language words. Therefore, it is difficult to ensure that the position I of word e is fixed, but it may appear near position i, in context [7]. Therefore, if the initial fixed position I is changed into a dynamic value so that it can slide within a certain range of the initial value, it will participate in the calculation of a posteriori probability when the target word appears within the limited range. This is the improved method based on position window on the basis of fixed position. Note that the position window is TR_k, A is the window size, which is a natural number. If the word appears within the position window, it is also considered that the word appears in the current translation hypothesis. Therefore, the posterior probability of a word can be determined by the sum of the posterior probability of the word at the position within the window. The calculation formula is as follows:

$$p_{i,t}\left(X \mid Y_1^J\right) = A\sum_{k=i-t}^{1+t} TR_k - p_i\left(e \mid f_1^J\right) \tag{2}$$

The statistics to be divided into equivalence classes according to frequency N_{c+1}, and then the statistics of equivalence classes with frequency plus one are used to estimate the frequency of the current class N_c. The calculation formula of using sparse representation method to estimate the occurrence times of C tuples in sparse representation model is as follows:

$$C^* = (C+1)\frac{N_{c+1}}{p_{i,t}\left(X \mid Y_1^J\right) - N_c} \tag{3}$$

Where λ_i is frequency of occurrence of specific n-tuples $f_i(x, y)$ frequency of occurrence of specific n+1 tuples; The number of p tuples whose frequency in m training set is C; m is the number of n+1 tuples with frequency n, m in the training set. Maximum entropy classifier is a generalization model of naive Bayesian classifier.

Its essential idea is to establish a continuous model for all known factors without considering any unknown factors. One of the main advantages of modeling using maximum entropy method is that different features can be easily added to the model [8]. The binary classifier sample is represented by (x_n, y_n) where X represents the feature sample. After the feature vector of the word is given, the formula for predicting the correctness and error of the word by using the maximum entropy model is shown in the formula.

$$p(y/x) = \frac{(x_n, y_n) + \exp\left(\sum_i p(n,m)\right)}{C^* \sum_y E\left(\mathrm{Re}\, xp\left(\sum_i \lambda_i f_i(x,y)\right)\right)} \tag{4}$$

The sparse representation model based on the part of speech statistics results can calculate the probability of each sentence. We assume that the probability value can effectively reflect whether a sentence is correct or not, or we assume that the probability value of the correct sentence is greater than that of the wrong sentence [9]. Therefore, judging whether there is a grammatical error in a sentence directly depends on the probability value of the sentence. It can be seen from the formula that the length of the sentence directly affects the probability value of the sentence. In order to reduce the impact caused by the length of the sentence, set the weight of each sentence $C(u)$. The specific method is as follows: convert any sentence with a part of speech sequence length of more than 6 according to the formula:

$$SL(S) = p(y/x) \sum_{n=6}^{\mathrm{len}(S)} \left(mn \times \sum_{u \in \mathrm{SubStr}(S,n)} \log(C(u)) \right) \tag{5}$$

In order to facilitate calculation and formal comparison, the weight of the sentence is normalized as follows:

$$SL(S) = \sum_{n=6}^{len(S)} \left(\frac{n \times \sum_{u \in SubStr(S,n)} \log(C(u))}{|Substr(S,n)|} \right) \tag{6}$$

$\mathrm{Substr}(S,n)$ is used to represent the $p_{\mathrm{lev}}\left(e, f_1^J, e_1^{'}\right)$ alignment relationship between the optimal translation $p_{\mathrm{lev}}\left(e, f_1^J, e_1^{'}\right)$ and other translation assumptions K in the sparse representation list, then for the word position I, the sentence posterior probability with Levenshtein alignment relationship can be given by the following formula:

$$p_{lev}\left(e/f_1^J, e_1^{'}\right) = KI - \frac{SL(S) + p_{lev}\left(e, f_1^J, e_1^{'}\right)}{\sum_{e^{'}} p_{lev}\left(e^{'}, f_1^J, e_1^{'}\right)} \tag{7}$$

Sentence fluency is usually measured by the sentences written by native English speakers, that is, it is considered that the sentences written by native English speakers are fluent. The fluency of sentences can be calculated through language model, such as formula.

$$H(x) = p_{\text{lev}}\left(e / f_1^J, e_1^{'}\right) - \frac{\sum_{i=1}^{|x|} \log P\left(x_i | x_{<i}\right)}{1 + H(x) - |x|} \tag{8}$$

In the codec structure, the encoder $H(x)$ is mainly represents the input word sequence as a middle semantic vector x, takes the middle semantic vector $P\left(x_i | x_{<i}\right)$ as the input in the decoder, and predicts the word at the current time in combination with its own generation sequence at the previous moment. The feature of this structure is that it can flexibly handle unequal input and output sequences [10]. Therefore, this model designs an automatic correction model for English text syntax errors based on the encoder and decoder structure of transformer model. The overall flow chart of the model is shown in figure 2.

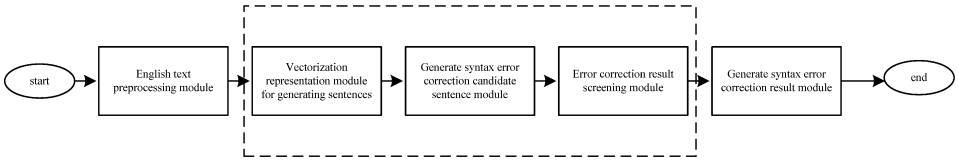

Figure 2. Processing flow chart of English text grammatical errors

Among them, the English grammar error detection and correction module based on sequence annotation includes preposition error detection and correction module and article error detection and correction module [11]. The English syntax error detection and correction module based on sparse representation includes encode module and decode module. After receiving the external request, first obtain the training data. The training data comes from two parts, one is the original training corpus, and the other is the accumulated user suggestion text. After the two parts of corpus are successfully obtained, the number of sentences in the new training corpus is counted. When the set new threshold is reached, the model training is started. Otherwise, the training exception is sent to the administrator for inspection [12]. Before the training, the corpus is regularized by using the corpus preprocessing script, and then the model is initialized and trained. This model is mainly composed of five modules, including English text preprocessing module, vectorization representation module of generating sentences, generating candidate sentences for grammatical error correction, error correction result screening module and generating grammatical error correction results module [13].

2.2 Error Evaluation Algorithm for English Text Image

The most commonly used evaluation algorithm for syntax error correction is the latest evaluation algorithm proposed by Max match by the University of Singapore, and it is also the evaluation method adopted in conli-2014. This method accurately compares the given error correction results with the reference answers, and gives the final model evaluation by comprehensively considering the correction rate and error correction rate [14]. Text grammatical errors are mostly manifested in the use of part of speech, tense or articles and prepositions in an English sentence. This paper makes statistics on conll's English grammar error detection and correction evaluation data in 2013 and

2014. The statistical proportion is shown in table 2.

Table 2. Statistics of English grammar error detection and correction evaluation tasks

Error type	Training set		Test set	
	Number (PCs.)	Proportion (%)	Number (PCs.)	Proportion (%)
article	6662	15.1	692	19.8
preposition	2412	5.5	315	9.1
noun	3782	8.5	395	12.2
Subject predicate consistency	1453	3.8	125	3.7
Verb form	1528	4.1	126	3.8
Five types	15832	35.8	1654	48.6
All types	45165	100.0	3571	100.0

There are many error types marked in the data, but the evaluation task is mainly aimed at five error types: Article error, preposition error, noun error, subject predicate consistency and verb form error [15]. From the statistical results, from the distribution of the five common types of English grammatical errors, the errors of articles and prepositions account for a high proportion; Moreover, the confusion set of preposition and article errors is relatively fixed. English grammar error detection and correction is designed into three modules.

The architecture of English grammar error detection and correction system includes text preprocessing module, English grammar error detection and correction module based on sequence annotation, English grammar error detection and correction module based on sparse representation, etc. When any step in corpus preprocessing, model initialization and model training is abnormal, a training exception notice will be sent to the administrator for repair. After the model training, according to the training results, use the evaluation script of the previous grammar error correction experiment to evaluate the model. If the error correction effect is improved, update the error correction model of the grammar error correction module, otherwise it will end directly.

2.3 Implementation of English Grammar Error Detection

There are two main non word error checking methods in English text: sparse representation analysis method and dictionary lookup method. Sparse representation analysis method is to find each n-ary string in the input string in the pre edited sparse representation table (N generally takes 2 or 3). N-ary strings that cannot be found or appear very frequently in the sparse representation table are considered to be possible spelling errors. Sparse representation analysis usually requires a dictionary or large-scale text corpus to edit the sparse representation table in advance. The dictionary lookup method mainly checks whether the input n-ary string is in the dictionary or acceptable vocabulary. If not, the input string will be marked as a misspelled word. The proofreading based on dictionary lookup method has high error checking accuracy and is a popular error detection technology at present. In the experiment, we use the dictionary search method to check non word errors.

Syntax error correction module is the core module, which mainly has three functions: data processing, model training and model error correction, of which model error correction is the core function. Data processing is responsible for cleaning and screening the original corpus, extracting effective text and carrying out structured processing to obtain regular text for later use. Model training, realize the syntax error correction algorithm, conduct model training and error correction effect evaluation

combined with error correction corpus, and save the trained model for testing and formal use. Model error correction: use the trained error correction model to correct the errors of sentences and output sentences that do not contain grammatical errors. The module will provide two thrift service interfaces: model training and model error correction. The former is responsible for receiving the request of model training and retraining and evaluating the algorithm model. The model error correction interface is responsible for syntax error correction of sentences and returns the error correction results. The syntax error correction module provides a model training interface for external trigger model retraining and updating to improve the effect of syntax error correction, which belongs to an important part of self updating.

3. Analysis of Experimental Results

The training data in the experiment mainly comes from 5607 wrong sentences and 5607 correct sentences corresponding to in the public data set. These sentences are trained by native English speakers to manually mark the grammatical errors in the article and correct each error. On the basis of these data sets, 11G data and 1150 wrong sentences and 1150 corresponding correct sentences in bilingual corpus are obtained from news corpus respectively. The error types and sentence expressions of sentences in bilingual corpus are similar to those in original corpus. A total of 1750 sentences were used in the experiment, which were provided by the public testing platform. The data set in this paper consists of three parts: training set, development set and test set. The training set adopts nucleus corpus, CLEC and icnale corpus. Nucleus is the official training corpus of the grammar error correction tasks conll-2013 and conll-2014, including 57151 parallel sentence pairs. The icnale corpus contains about 1.3 m tokens, all of which are from Asian ESL learners. The CLEC corpus is fully applicable to the training of GEC developed for Chinese English learners. The CLEC corpus is divided according to the composition topics, and 1000 compositions are randomly selected from five topics as the test set to test the applicability of this model to Chinese ESL learners, Take the composition of all other topics as the training set and express it with "allexcept5 titles". Conll-2014 test set and jfleg test set are used in the test. In order to be more clear and intuitive and facilitate the subsequent formula description, the above sample types are summarized in the form of matrix, as shown in the table 3 below.

Table 3. Matrix representation of sample types

Model judgment Manual marking	Positive sample	Negative sample
Positive sample	WI	QM
Negative sample	QI	WM

According to the proportion of different types of samples, the performance of the error correction model is evaluated. The calculation method of evaluation indicators will be further introduced below. In this study, there are 10071 sentences in the training set, including 24797 sentences with grammatical errors, and 9033 sentences in the test set, including 3316 sentences with grammatical errors. For example, the research published by Alibaba in 2017 found that adding sentences with correct grammar can improve the model results to a certain extent. Therefore, in this study, the author added a higher proportion of correct sentences in the training shown in Table 4.

Table 4. Data set distribution statistics

	All sentences	Grammatically correct sentences	Sentences with grammatical errors
Training set	100078(100%)	75985(75.35%)	24758(25.52%)
Test set	9124(100%)	4785(52.22%)	4512(49.85%)

Each sentence with grammatical errors contains at least one grammatical error. The proportion of the training set is slightly higher than the conventional proportion because it contains a higher proportion of grammatically correct sentences in the training set. Obfuscate the set to improve the accuracy of rule-based syntax error correction. In the classification method, text syntax error correction is regarded as a multi classification problem, and a confusion set is specified for a given error type. The features used include part of speech tags and dependent sentences. The features in this experiment refer to the combination of words and parts of speech. For specific error types, the error correction task is regarded as a classification task that can learn the syntax representation from a large number of local text data. Comparing and verifying the accuracy of common text classification algorithms tf-df and naive Bayesian algorithm, it is found that the classification effect of these two algorithms is not good in the application scenario of this experiment. Through the experiment, it is found that when the confusion set is expanded, the SVM classification effect is better, so we use SM algorithm to define the confusion set for the specified error type, the error correction accuracy has been significantly improved, as shown in the Figure 3 below.

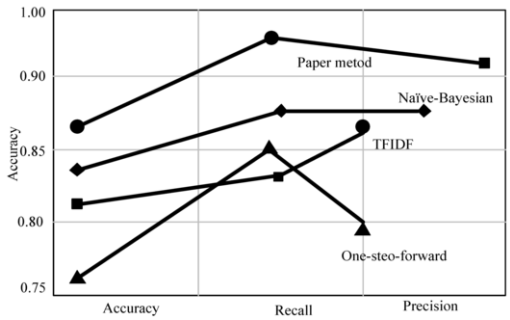

Figure 3. Comparison and detection results of syntax error detection accuracy of different methods

In this study, the author added a higher proportion of correct sentences in the training and. In addition, we made statistics on the types of grammatical errors in these 1000 English compositions. In the original annotation of CLEC corpus, the classification of grammatical errors is detailed, but the number is too large. Therefore, we corresponding and summarized the types of grammatical errors marked in CLEC according to the classification standard of conli-2014, The number of mark errors, model detection errors and correct correction errors of each syntax error type are calculated and displayed. In order to be more intuitive, we calculate the accuracy rate and recall rate of each syntax error type according to the error correction results, which are expressed in the form of a graph, as shown in the figure 4 below.

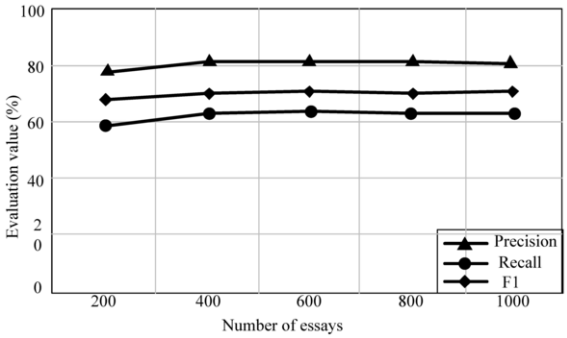

Figure 4. Syntax error correction evaluation values of this model under different number of articles.

The experimental results of CRF model with word and part of speech as features and with dependent syntax tree structure as extended features based on word and part of speech features are shown respectively in Table t and Table 6.

Table 5. Experimental results of CRF classification based on word and part of speech

performance index	Accuracy	Precision	Recall	F-Score
Detection layer	0.5125	0.5685	0.1089	0.1852
Identification layer	0.4825	0.3652	0.0512	0.0829

Table 6. Experimental results of CRF classification with dependency syntactic structure features

performance index	Accuracy	Precision	Recall	F-Score
Detection layer	0.5365	0.5581	0.3562	0.4328
Identification layer	0.4825	0.4125	0.1893	0.2615

The experimental results show that the detection results based on sparse representation can achieve high accuracy, as depicted in figure 5. The experiments show that the model can correct the errors of articles and qualifiers Noun singular and plural errors (verb form errors and modal verbs) have good effects, especially for subject predicate consistency errors and verb deletion. This is mainly due to the combination of transformer and bi-gru, which increases the scope of feature extraction. At the same time, the improved error detection method can obtain more accurate reasoning results during decoding.

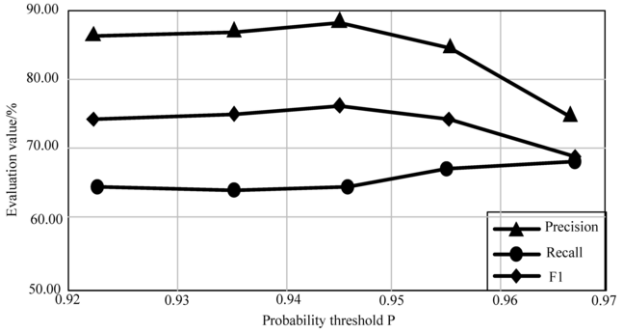

Figure 5. Influence of different probability thresholds on error detection values.

In order to obtain the optimal probability threshold of the improved error detection,

the selection of the threshold is tested. The accuracy rate, recall rate and F1 value of the model are investigated, and the threshold P most suitable for the model is selected according to F. the experimental results are shown in the figure. It can be seen from the experimental results that when the value of P is 0.945, it can maximize the retention of high accuracy while considering more candidate results. Therefore, the formal method in this paper has high detection accuracy and fully meets the research requirements.

4. Concluding Remarks

Based on sparse representation, the maximum entropy classifier is used to recognize three typical WPP features, linguistic features, source word features, three different English text image features and linguistic vocabulary features, optimize the grammar error detection algorithm and improve the grammar error detection process. The experimental results show that the grammar error detection method based on sparse representation is effective in detecting translation errors.

References

[1] Liu Shuai, Liu Dongye, Muhammad Khan, Ding Weiping. Effective Template Update Mechanism in Visual Tracking with Background Clutter. Neurocomputing, 2021, 458: 615-625.
[2] Liu Shuai, Wang Shuai, X Liuinyu, Amir H. , Gandomi, Mahmoud Daneshmand, Khan Muhammad, Victor Hugo C. De Albuquerque, Human Memory Update Strategy: A Multi-Layer Template Update Mechanism for Remote Visual Monitoring, IEEE Transactions on Multimedia, 2021, 23:2188-2198
[3] Liu S, He T, Dai J., A Survey of CRF Algorithm Based Knowledge Extraction of Elementary Mathematics in Chinese. Mobile Networks & Applications, 2021, 1-13.
[4] Abeyrathna KD, Granmo OC, Goodwin M. Adaptive Sparse Representation of Continuous Input for Tsetlin Machines Based on Stochastic Searching on the Line. Electronics, 2021, 10(17):2107.
[5] Yz A, Yong M, Xda B, Hao LC, XMA B, JMA B. Locality-constrained sparse representation for hyperspectral image classification. Information Sciences, 2021, 546(3):858-870.
[6] Ghasemian N, Shah-Hosseini R. Hyperspectral multiple-change detection framework based on sparse representation and support vector data description algorithms. Journal of Applied Remote Sensing, 2020, 14(1):1.
[7] Tarawneh A S, Celik C, Hassanat A B, Chetverikov D. Detailed investigation of deep features with sparse representation and dimensionality reduction in CBIR: A comparative study. Intelligent Data Analysis, 2020, 24(1):47-68.
[8] Aydin I, Kaner S. A New Hybrid Diagnosis of Bearing Faults Based on Time-Frequency Images and Sparse Representation. Traitement du Signal, 2020, 37(6):907-918.
[9] Sreelakshmi K, Premjith B, Soman K P. Detection of Hate Speech Text in Hindi-English Code-mixed Data. Procedia Computer Science, 2020, 171(3):737-744.
[10] Merseedi KJ. 2020, CREATE A NEW FUNCTION TO CONVERT NUMBERS INTO TEXT FORM BY MS-EXCEL IN THREE LANGUAGES (ENGLISH, KURDISH, AND ARABIC). Technology Reports of Kansai University, 62(3):511-517.
[11] Rahman M M, Y Watanobe, Nakamura K. A Neural Network Based Intelligent Support Model for Program Code Completion. Scientific Programming, 2020, 1:1-18.
[12] Rm A, Dr B, Om B, RM A. Decompiled APK based malicious code classification. Future Generation Computer Systems, 2020, 110(2):135-147.
[13] Kwan YY, Tan CK. THE USE OF KAHOOT! IN IMPROVING UNDERGRADUATE STUDENTS' ENGLISH GRAMMAR ACHIEVEMENT: A CASE STUDY IN A PUBLIC UNIVERSITY IN SABAH. Solid State Technology, 2020, 63(1):1539-1556.
[14] J ATD Silva. Outsourced English revision, editing, publication consultation, and integrity services should be acknowledged in an academic paper. Journal of Nanoparticle Research, 2021, 23(4): 81.
[15] Silva L, Neto E, Francisco R, Barbosa JLV, Leithardt VRQ. ULearnEnglish: An Open Ubiquitous System for Assisting in Learning English Vocabulary. Electronics, 2021, 10(14):1692.