IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 1, 2019, accepted November 25, 2019, date of publication December 4, 2019,
date of current version December 23, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2957582

An Automated Grader for Chinese Essay Combining Shallow and Deep Semantic Attributes

YIQIN YANG[1], LI XIA[2,3], (Senior Member, IEEE), AND QIANCHUAN ZHAO[1], (Senior Member, IEEE)

[1]Center for Intelligent and Networked Systems, Department of Automation, Tsinghua University, Beijing 100084, China
[2]Business School, Sun Yat-Sen University, Guangzhou 510275, China
[3]Guangdong Province Key Laboratory of Computational Science, Sun Yat-Sen University, Guangzhou 510275, China

Corresponding author: Li Xia (xial@tsinghua.edu.cn; xiali5@sysu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2016YFB0901900 and Grant 2017YFC0704100, and in part by the National Natural Science Foundation of China under Grant 61573206, Grant 11931018, and Grant U1301254.

ABSTRACT Writing is a pivotal part of the language exam, which is considered as a useful tool to accurately reflect students' language competence. As Chinese language tests become popular, manual grading becomes a heavy and expensive task for language test organizers. In the past years, there is a large volume of research about the automated English evaluation systems. Nevertheless, since the Chinese text has more complex grammar and structure, much fewer studies have been investigated on automated Chinese evaluation systems. In this paper, we propose an automated Chinese essay evaluation system called AGCE (Automated Grader for Chinese Essay), which combines shallow and deep semantic attributes of essays. We implement and train our AGCE system on a Chinese essay dataset, which is created by ourselves based on more than 1000 student essays from a Chinese primary school. Experimental results indicate that our AGCE system achieves the quadratic weighted Kappa of 0.7590 on a small dataset, which is of higher grading accuracy compared with other four popular neural network methods trained on large-scale datasets. In addition, our AGCE system can provide constructive feedback about Chinese writing, such as misspelling feedback and grammatical feedback about writers' essays, which is helpful to improve their writing capability.

INDEX TERMS Automated Chinese essay evaluation, AGCE, natural language processing, semantic attributes, semantic feedback.

I. INTRODUCTION

Writing is a creative process including analysis, thinking, and speculation. It can accurately reflect the writers' level of vocabulary, grammar and organization competence. For this reason, writing plays a pivotal role in the language exam. Regarding grading writing, manual grading is extensively used. However, it is time-consuming and inefficient. Automated essay evaluation (AEE) provides a solution to this problem. It can not only score writers' essays but also provide some constructive feedback, which could save teachers' time without lowering the quality.

Page carried out the work of the AEE field and proposed the first automated essay scoring (AES) system in 1966 [1]. This system merely used the basic attributes to describe the quality of an essay. By the 1990s, the development of natural language processing (NLP) encouraged researchers to employ new techniques to extract deep attributes in essays. Since AEE systems enable students to receive constructive feedback about their essays based on these deep attributes, the AEE systems have gradually replaced AES systems in the last decade and are widely adopted in educational settings [2]. Current AEE systems cooperate with human graders to accomplish tasks in some language exams such as Graduate Record Examination (GRE), Test Of English as a Foreign Language (TOEFL), and American College Testing (ACT) [3]. If scores between AEE systems and human graders differ by more than a certain level, another human grader will give the final score.

The existing literature on AEE systems is extensive and focuses particularly on English text. Some AEE systems, such as the IntelliMetric system [4], could analyze a text in multiple languages. However, there are relatively few historical studies in the area of automated Chinese essay evaluation systems. In addition, some AEE systems, such as the PEG system, are easily cheated [5] because they primarily focus

The associate editor coordinating the review of this manuscript and approving it for publication was Zhenliang Zhang.

on surface attributes rather than semantic attributes [6]. Other AEE systems either ignore the timing sequence information in an essay or just add new attributes based on the Intelligent Essay Assessor (IEA) system to improve accuracy [7]. Since most AEE systems are highly reliant on training datasets, it cannot use its own experience like a teacher to analyze a new genre essay outside of the training datasets [8].

Compared with the English text, Chinese regularly uses words that are only grammatical (i.e. they express relationships between concepts) or even connotational (they imply additional information) rather than notional (expressing concepts) [9]. Therefore, Chinese text has much more complex grammar and structure, which limits the development of automated Chinese essay evaluation systems. The contribution of this paper is that we propose and develop an extended of AEE system for Chinese text, which combines shallow and deep semantic attributes of an essay. The shallow attributes could directly describe the quality of the essay, which include the number of words and sentences. Datasets of Chinese essays are much less available, compared with datasets of English essays. We collect more than 1000 Chinese student essays from a Chinese primary school and set up a standard dataset for Chinese essays. Based on this dataset, we implement our Chinese AEE system and we name it Automated Grader for Chinese Essays (AGCE). The deep semantic attributes including spelling mistakes and grammatical errors detected based on the timing sequence model. As an AEE system, our AGCE system can provide constructive feedback, and accurately reproduce the human graders' scores. As there is a recent surge of interest in deep learning, Taghipour reported a method based on neural networks for the AES task. These neural networks were trained on the dataset provided by the Automated Student Assessment Prize (ASAP) competition and outperformed the Enhanced AI Scoring Engine (EASE) system by 5.6% in terms of quadratic weighted Kappa. The EASE system is the best opensource system participating in the ASAP competition, which is ranked third among all 154 teams. We compare our system with four neural networks trained by Taghipour, and our AGCE system achieves higher grading accuracy on a small dataset.

The remainder of this paper is outlined as follows. In a Section II, we describe the related work of AEE systems. In Section III, we give a framework of our AGCE system. We propose related methods to extract shallow and deep semantic attributes in Section IV and Section V, respectively. In Section VI, we present the implementation and evaluation of the proposed system. We demonstrate the results of our AGCE system in Section VII. Finally, we conclude this paper in Section VIII.

II. RELATED WORK

A considerable amount of literature has been published on AEE systems for English text. In 1966, Page proposed the Project Essay Grader (PEG) system [1]. He measured multiple attributes including fluency, grammar, and punctuation, and used a linear regression method to predict the essay scores [10]. In 1996, Latent Semantic Analysis (LSA) was created by Foltz [11]. This method represents text as a matrix, which is used to measure the semantic similarities between words. However, it is difficult to determine the number of matrix dimensions and the semantic information needs to be provided by a large corpus of texts [12]. In 2004, Person Knowledge Technologies presented the IEA system based on LSA. Compared with the PEG system, the IEA system created three sources, which include pre-scored essays of other students, high-score essays evaluated by experts and an unscored set of essays. The IEA system would compare unscored essays with pre-scored essays on the same topic to obtain unscored essays scores. Since the IEA system represents these essays as a matrix using LSA, the IEA system just requires 100 pre-scored essays [13].

From the 1990s to 2010s, the electronic essay rater (E-rater) system was developed by the Educational Testing Service (ETS), and it was applied in GMAT and TOFEL. In contrast to the IEA system, the E-rater system adopted copy-edited text sources to build its corpus and model. This approach consists of three modules including syntactic module, discourse module, and topical-analysis module [14]. The E-rater system employed these modules to obtain a mathematical representation of an essay and saved it into the training datasets. For this reason, the accuracy of the E-rater system highly depends on training datasets. In 2006, the IntelliMetric system was proposed by Vantage Learning [4]. The IntelliMetric system evaluates over 300 attributes by classifying these attributes into five broad categories, which include focus and unity, organization, development and elaboration, sentence structure, mechanics and conventions [15], [16]. Since the IntelliMetric system could process text in multiple languages and provide instructive feedback, the IntelliMetric system becomes a standardized assessment model [17]. In 2002, the Betsy system was developed as a research tool by Ruder and Liang [18]. On the basis of the Bayesian theorem, the Betsy system adopted the best attributes from the E-rater system, and it was simple to apply the Betsy system in short essays and various genres essays [19]. However, the research on the Betsy system is limited. The Betsy system is merely suitable for language research, not for students.

In addition to the above attributes provided by AEE systems, Persing proposed methods to analyze other attributes of an essay. Persing asked human annotators to score each of the 1000 argumentative essays, which were selected from the ICLE corpus along the argument strength dimension. This method represents predicting the argument strength score of an essay as a regression problem [20]. Persing employed similar methods to analyze the thesis clarity dimension [21], the prompt adherence dimension [22], and the modeling organization [23]. The concrete analysis of attributes is instructive for designing AEE systems [24].

Compared with AEE systems for English text, there is a much smaller body of literature that is concerned with automated Chinese essay evaluation systems. In 2010, Peng attempted to adopt several vector space models and

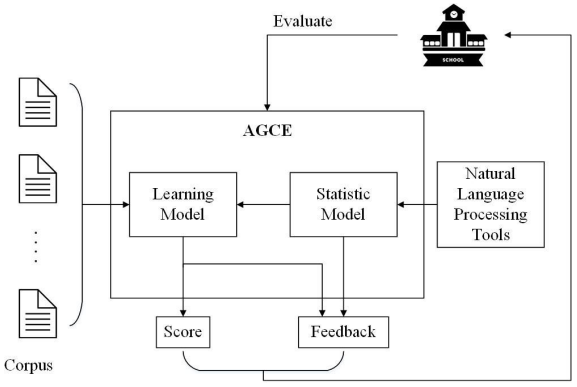


FIGURE 1. The framework of our AGCE system.

added some statistical surface features for Chinese text [25]. In 2016, Hao proposed the SCESS system based on Weighted Finite State Automata (WFSA) and used Incremental Latent Semantic Analysis (ILSA) to deal with a large number of essays [26]. The SCESS system constructed a WFSA to perform text pre-processing based on an N-gram language model and used ILSA to perform automated essay scoring. These studies on Chinese essay scoring are not systematic, compared with the English AEE systems. Moreover, the Chinese essay dataset is also not available, compared with the English essay datasets. Chinese AEE systems need more research attention, which is also the main objective of this paper.

III. THE PROPOSED SYSTEM

In this paper, we propose a so-called AGCE system for Chinese essay scoring, which consists of a statistic model and a learning model, as illustrated in Fig. 1. The statistic model could extract shallow attributes based on natural language processing tools. The learning model is used to extract deep semantic attributes and analyze these feature vectors.

In this paper, the shallow attributes include the number of words, pinyin, sentences and metaphors, the length of sentences and the number of different levels of vocabulary. These shallow attributes extracted by the statistic model can directly reflect the quality of an essay. In the last years, Gutierrez proposed that AEE systems should recognize some certain type of errors including syntactic errors, which plays a pivotal role in improving students' writing level. In this case, our AGCE system not only achieves accurate grading but also provides instructive feedback. The learning model would analyze feature vectors and extract deep semantic attributes including misspelling attributes and grammar attributes.

The input of the AGCE system is the manuscript photo of students, as illustrated in Fig. 2. It is necessary to apply OCR technology to identify Chinese characters in the photo and clean the results. In this process, we would preserve spelling mistakes, pinyin and correct the results manually. Since the aesthetics of the manuscript photo would have the score, we remove manuscripts with poor handwriting, and then the score of an essay is merely related to its content, not for penmanship.

FIGURE 2. The manuscript photos of students essays in datasets.

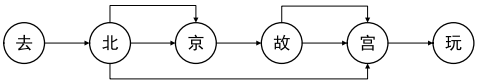


FIGURE 3. The directed acyclic graph of an example.

The conversion from input to feature vector is

$$s_1 = function_1(w^{(q)}), \quad q = 1, 2, \cdots, Q \tag{1}$$

$$s_2 = function_2(w^{(q)}), \quad q = 1, 2, \cdots, Q \tag{2}$$

$$s = \{s_1, s_2\} \tag{3}$$

where w is the input of the AGCE system, Q is the number of samples in datasets and s is the feature vector, $function_1$ and $function_2$ are employed to extract shallow and deep semantic attributes, respectively. The relationship between the output t and the feature vector is

$$\left\{s^{(q)} : t^{(q)}\right\}, \quad q = 1, 2, \cdots, Q \tag{4}$$

IV. SHALLOW ATTRIBUTES EXTRACTION

A. PROBABILISTIC SEGMENTATION MODEL

Each essay has a reasonable sequence of words. English essays adopt spaces as a natural separation symbol. However, Chinese essays have more than one character in a phrase without spaces, which results in more difficulties in Chinese word segmentation. The probabilistic segmentation model provides a solution to this problem. We first construct a prefix dictionary and directed acyclic graph, as illustrated in Fig. 3.

Each word in the directed acyclic graph has its weight, which is the word frequency in datasets. The unrecognized word can be considered as a single phrase. In this way, we can represent the probabilistic segmentation model as a Hidden Markov Model (HMM) [27]

$$P(q_t = s_j | q_{t-1} = s_i, q_{t-2} = s_k \cdots) = P(q_t = s_j | q_{t-1} = s_i), \tag{5}$$

where $S = \{s_1, s_2, \cdots\}$ is the words to be processed, the $t - th$ word is merely related to the previous word. Compared with Markov models, HMM contains observable states and hidden states, as shown in Fig. 4. The hidden states of the probabilistic segmentation model include S, B, M, E, which respectively represent a single phrase, the beginning, middle and end of a phrase. The observable states are Chinese characters.

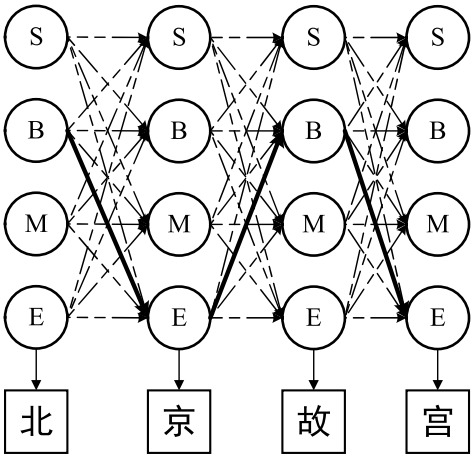


FIGURE 4. The probabilistic segmentation model, where the hidden states are segmentation symbols and the observable states are Chinese characters.

We would give a mathematical description as follows. The maximum probability of $t-th$ word and $i-th$ hidden state in a single path is defined

$$\delta_t(i) = max\{P(i_t = i, i_{t-1}, \cdots, i_1, o_t, \cdots, o_1|\lambda)\}, i = 1, 2, \cdots, N \quad (6)$$

where λ is the parameter of probabilistic segmentation model, o_t is the observable state, a is the transition matrix and b is the observable probabilistic matrix. Regarding the $t-th$ word, the $(t-1)th$ node of the $i-th$ hidden state in a maximum probabilistic path is defined

$$\psi_t(i) = argmax_{1\le j\le N}[\delta_{t-1}(j)a_{ji}] \quad (7)$$

The $\delta_{t+1}(i)$ can be calculated by

$$\delta_{t+1}(i) = max[\delta_t(i)a_{ji}]b_i(o_{t+1}) \quad (8)$$

The termination state is

$$P^* = max_{1\le i\le N}\delta_T(i) \quad (9)$$

$$i_T^* = argmax_{1\le i\le N}[\delta_T(i)] \quad (10)$$

We can backtrack the optimal path of $T-1, T-2, \cdots, 1-th$ words

$$i_t^* = \psi_{t+1}[i_{t+1}^*] \quad (11)$$

The optimal path, which means the optimal sequence of words is

$$I^* = (i_1^*, i_2^*, \cdots, i_T^*) \quad (12)$$

B. LEXICAL ATTRIBUTES

The lexical attributes can directly reflect the quality of an essay [28]. The IntelliMetric system has extracted hundreds of attributes from datasets. For this reason, we adopt the three most effective attributes from the IntelliMetric system, which include the number of words and sentences, and the length of sentences. Generally, when the three attributes are greater, writers would be considered having stronger language competence. However, compared with English text, word order in Chinese may be considered as largely inflexible, and syntactic groups are positioned in a sentence following strict rules [9]. For this reason, Chinese text has a more complex description, and the three attributes may have limited competence. To deal with this problem, we add other attributes based on Chinese text to evaluate essays.

像 似 如同 如 似的 好比 好像 犹如

FIGURE 5. The metaphorical relationships from textbooks of Chinese primary schools.

TABLE 1. Corpora of different grades.

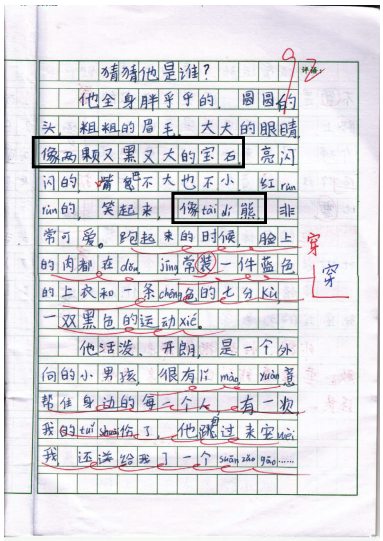
Grade	one	two	three	four	five	six
Size of corpus	174	536	1132	1737	2172	2655

In Chinese essays, metaphors are employed to replace abstract things with simple, concrete and vivid things. In general, the number of metaphors can reflect the essay's description. The basic structure of metaphors can be divided into three parts: a metaphorical thing, a metaphorical relationship and an analogical thing. In Fig. 6, there are two metaphors and they have the same metaphorical relationship. In this work, we can employ metaphorical relationships to detect metaphors in an essay. The metaphorical relationships are illustrated in Fig. 5.

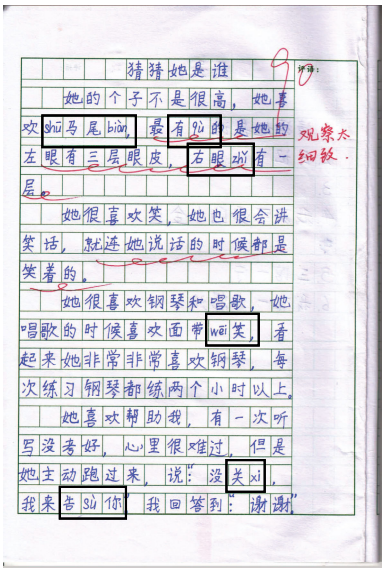
The Chinese orthography does not map into the sound system altogether, in contrast to the English alphabet, which maps (at least in large part) into the level of phonemes [29]. Due to the complexity of Chinese spelling, some Chinese characters cannot be spelled correctly by beginners and therefore be represented with pinyin, as illustrated in Fig. 6. In Chinese text, Each word has its pinyin, which can determine the word pronunciation. Pinyin is similar to the international phonetic alphabet in English text. Essays with much pinyin usually mean that writers might have weaker language competence.

English text is composed of different levels of words, and the high-level words can heavily improve the quality of an essay. In addition to the above attributes, the quality of vocabulary in an essay can be considered as a crucial part of the content. Our AGCE system serves students in primary school, and these students gain knowledge from textbooks. For this reason, we preprocess official textbooks of the primary school with the probabilistic segmentation model in Section IV, and construct corpora of different grades, as presented in Table 1.

Since there is less content in the grade-one textbook compared with the grade-six textbook, the number of grade-one vocabulary in the corpora is small, while the number of grade-six vocabulary is large. In addition, different levels of verbs and idioms are respectively included in these corpora, and we do not consider them as attributes. The shallow attributes we have counted are presented in Table 2.



(a) Chinese essay including metaphors



(b) Chinese essay including pinyin

FIGURE 6. Chinese essays including metaphors and pinyin, which are circled by a black box.

TABLE 2. Shallow attributes.

	Lexical attribute		Words of different grades
1	number of words	6	grade-one of vocabulary
2	number of sentences	7	grade-two of vocabulary
3	length of sentences	8	grade-three of vocabulary
4	number of metaphors	9	grade-four of vocabulary
5	number of pinyin	10	grade-five of vocabulary
		11	grade-six of vocabulary

V. DEEP SEMANTIC ATTRIBUTES EXTRACTION

In the past years, many researchers have debated that AEE systems not only need to accurately reproduce the human graders but also recognize some certain types of errors, such as spelling and grammar errors [30], [31]. In this paper, the learning model in our AGCE system can automatically recognize these semantic errors. The logic framework of the learning model is illustrated in Fig. 7 and described in Algorithm 1. We would give a detailed description of the learning model as follows.

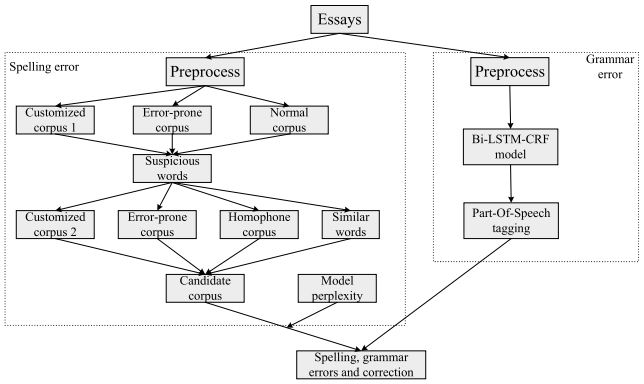


FIGURE 7. The logic framework of the learning model in our AGCE system.

Algorithm 1 Automated Error Detection (AED) System

Input:
spelling error recognition corpora, correction corpora, ungraded essay

Output:
spelling error and correction, grammar error

```
Preprocessing
for words in text do
    if words not in spelling error recognition corpora then
        Add to suspicious corpus
    end if
end for
for words in suspicious corpus do
    Select candidate words in correction corpora
    Calculate language perplexity (PP)
    spelling error ⇐ suspicious words
    correction ⇐ candidate words with minimal PP
end for
Words embedding
Tagging sequence based on Bi-LSTM-CRF model
grammar error ⇐ words with R, M, S, W
```

TABLE 3. Spelling error recognition corpora.

Corpus	Number
customized corpus 1	177
error-prone corpus	759
normal corpus	584429

A. SPELLING ERROR

We initially employ the probabilistic segmentation model in Section IV to preprocess an essay. Then we remove the punctuation, and construct spelling error recognition corpora, which include customized corpus 1, error-prone corpus and normal corpus. These corpora are presented in Table 3.

TABLE 4. Correction corpora.

Corpus	Number
customized corpus 2	3502
error-prone corpus	759
homophone corpus	3431
similar words	1664

In Chinese text, some words do not follow grammatical logic. In order to address this limitation, we count the commonplace names, person names and item names in the primary school textbooks as the customized corpus 1. If a word in an essay is included in the customized corpus 1, the word will be considered correct. We also count idioms with high misspelling frequency and adopt the erroneous forms of these idioms to construct an error-prone corpus. Words included in the error-prone corpus will be considered wrong, and then be added to the set of suspicious words. Since the newspaper text contains numerous customary words, we count words in newspaper text to construct the normal corpus. The words included in this corpus would be considered correct. Unlike the toolkit ''pycorrector'', which adopts the N-gram model [32] to detect misspelling error, our AGCE system directly adds the words outside these corpora into the set of suspicious words.

Then we construct correction corpora, which include customized corpus 2, error-prone corpus, homophone corpus and similar words. These corpora are presented in Table 4.

In Chinese essays, some words have limited error forms. For this reason, we count the words of high misspelling frequency and employ their corrective forms to construct the customized corpus 2 and error-prone corpus. If suspicious words are included in the two corpora, the corrective forms will be added to the candidate corpus. In addition, almost every Chinese character has its homophone, and the form of misspelling error is typically its homophone or similar words. In this case, we count the homophones and similar words to respectively construct the homophone corpus and the similar words corpus. If suspicious words are included in the two corpora, its homophone or similar words will be added to the candidate corpus.

We define a sentence based on the candidate corpus, $S = \{w_1, w_2, \cdots, w_n\}$, and the length of this sentence is n. The language perplexity is calculated by

$$PP(S) = \sqrt[n]{\prod_{i=1}^{n} \frac{1}{p(w_i|w_{i-1})}} \tag{13}$$

The language perplexity model is trained based on newspaper corpus. The words with the least language perplexity will be adopted as the final corrective result.

B. GRAMMAR ERROR

Unlike the misspelling recognition in Section V, we first vectorize an essay. The traditional word vector method is one-hot, which represents a word with a unique value in one dimension. Since the number of words in an essay determines the dimension of the word vector matrix, this matrix has a strong sparsity. In 2003, Bengio adopted neural networks to learn word vector matrix, which maps a high-dimensional sparse word vector to a low-dimensional representation space [33]. In 2013, Mikolov highly improved the neural network model and proposed a word vector training tool, Word2Vec [34]. The Word2Vec includes Continuous Bag-of-Words Model (CBOW) and Skip-gram model, which is illustrated in Fig. 8. We illustrate the concrete technological analysis as follows.

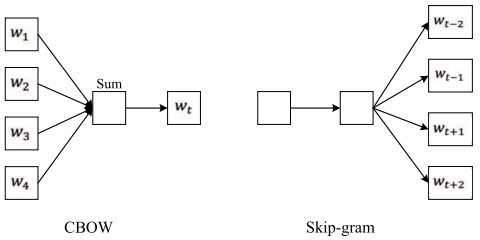


FIGURE 8. The logic framework of the Word2Vec model.

We randomly select a central word w_0. The context information of w_0 is $2c$, which is defined as $context(w_0)$ and positive sample. Next, using Negative Sampling, we can obtain neg negative samples, which are defined as $w_i, i = 1, 2, \cdots, neg$. This process converts multi-class classification problem into binary classification problem. The positive sample is calculated by

$$P(context(w_0), w_i) = \sigma(x_{w_0}^T \theta^{w_i}), \quad y_i = 0, \quad i = 1, 2, \cdots, neg, \tag{14}$$

where the weight of classifier is θ^{w_i} and σ is the activation function, sigmoid. The loss function is

$$L = \sum_{i=0}^{neg} y_i log(\sigma(x_{w_0}^T \theta^{w_i})) + (1 - y_i) log(1 - \sigma(x_{w_0}^T \theta^{w_i})). \tag{15}$$

The gradient of θ^{w_i} is

$$\frac{\partial L}{\partial \theta^{w_i}} = (y_i - \sigma(x_{w_0}^T \theta^{w_i}))x_{w_0}. \tag{16}$$

The gradient of x_{w_0} is

$$\frac{\partial L}{\partial x^{w_0}} = \sum_{i=0}^{neg} (y_i - \sigma(x_{w_0}^T \theta^{w_i}))\theta^{w_i}. \tag{17}$$

The word vector is obtained by iteratively calculating x_{w_0} and θ^{w_i} based on (16) and (17).

Then we take the trained word vector as the input of the Bi-LSTM-CRF model. The output of this model is a tagging sequence, which is composed of the part-of-speech tagging sequence and the grammatical error tagging sequence. The grammatical error tagging sequence has four types: redundant words (R), missing words (M), wrong words (S) and unordered words (W). The Bi-LSTM-CRF model is illustrated in Fig. 9 and we describe this model in detail as follows.

In 1997, Schmidhuber proposed the Long Short-Term Memory model (LSTM), which has forget gate, input gate and output gate. We define the state of LSTM cell as c,

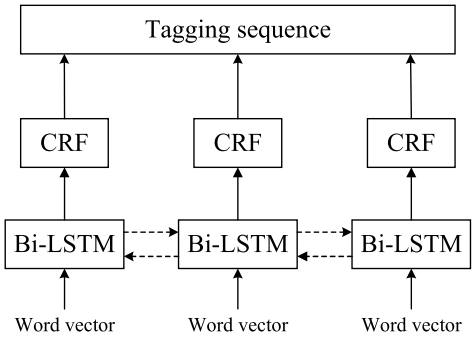


FIGURE 9. The logic framework of the Bi-LSTM-CRF model.

the output as a and the weight as w [35]. The forget gate is used to selectively forget the previous cell's output and state

$$f_t = \sigma(w_f \cdot [a_{t-1}, c_t] + b_f). \tag{18}$$

Then the input gate determines what the new information is stored in the cell

$$u_t = \sigma(w_u \cdot [a_{t-1}, c_t] + b_u), \tag{19}$$
$$\tilde{c_t} = tanh(w_c \cdot [a_{t-1}, c_t] + b_c), \tag{20}$$

where the outputs of the sigmoid layer are updated values, and the outputs of the tanh layer are candidate vectors. When the cell's state is updated, we remove some information and add new information

$$c_t = f_t \cdot c_{t-1} + u_t \cdot \tilde{c_t}. \tag{21}$$

Eventually, the output of LSTM cell is determined by the output gate

$$o_t = \sigma(w_0[a_{t-1}, w_t] + b_0), \tag{22}$$
$$a_t = o_t \cdot tanh(c_t). \tag{23}$$

The conditional random field (CRF) is employed to globally optimize the output of LSTM. The parameterized form of CRF is

$$P(y|x) = \frac{1}{Z(x)} exp(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)), \tag{24}$$

$$Z(x) = \sum_{y} exp(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} \mu_l s_l(y_i, x, i)), \tag{25}$$

where y_i is the output of LSTM, t_k, s_l are the characteristic function, λ_k, μ_l are the corresponding weight and $Z(x)$ is the normalization factor. We define that the number of transfer feature is K_1 and the number of state feature is K_2

$$f_k(y_{i-1}, y_i, x, i) = \begin{cases} t_k(y_{i-1}, y_i, x, i), & \text{if } k = 1, 2, \cdots, K_1 \\ s_l(y_i, x, i), & \text{if } k = K_1 + l; \\ & l = 1, 2, \cdots, K_2 \end{cases} \tag{26}$$

Then the transfer and state features are summed at each location

$$f_k(y, x) = \sum_{i=1}^{n} f_k(y_{i-1}, y_i, x, i), \tag{27}$$

where $k = 1, 2, \cdots, K$. We could represent $f_k(y, x)$ with w_k

$$w_k = \begin{cases} \lambda_k, & \text{if } k = 1, 2, \cdots, K_1 \\ \mu_l, & \text{if } k = K_1 + l; l = 1, 2, \cdots, K_2. \end{cases} \tag{28}$$

The (24) and (25) can be simplified as

$$P(y|x) = \frac{1}{Z(x)} exp(\sum_{k=1}^{K} w_k f_k(y, x)), \tag{29}$$

$$Z(x) = \sum_{y} exp(\sum_{k=1}^{K} w_k f_k(y, x)). \tag{30}$$

Since the empirical probability distribution based on dataset is $\tilde{P}(X, Y)$ and the parameter of CRF model is $w = (w_1, w_2, \cdots, w_k)^T$, the log likelihood function of training dataset could be represented as

$$L(w) = L_{\tilde{P}}(P_w) = log \prod_{x,y} P_w(y|x)^{\tilde{P}x,y} = \sum_{x,y} \tilde{P}(x, y) log P_w(y|x). \tag{31}$$

If P_w satisfies (29) and (30), the log likelihood function could be calculated by

$$L(w) = \sum_{j=1}^{N} \sum_{k=1}^{K} w_k f_k(y_j, x_j) - \sum_{j=1}^{N} log Z_w(x_j). \tag{32}$$

The transfer matrix t_k is

$$E_{\tilde{P}}[t_k] = \sum_{x,y} \tilde{P}(x) P(y|x) \sum_{i=1}^{n+1} t_k(y_{i-1}, y_i, x, i) exp(\delta_k T(x, y)), \tag{33}$$

where $\delta = (\delta_1, \delta_2, \cdots, \delta_K)^T$ is vector increment. The state matrix s_l is

$$E_{\tilde{P}}[s_l] = \sum_{x,y} \tilde{P}(x) P(y|x) \sum_{i=1}^{n} s_l(y_i, x, i) exp(\delta_{K_1+l} T(x, y)), \tag{34}$$

where $T(x, y)$ is the sum of all features in datasets

$$T(x, y) = \sum_{k=1}^{K} \sum_{i=1}^{n+1} f_k(y_{i=1}, y_i, x, i). \tag{35}$$

The δ_k and w_k can be updated by calculating (34) and (35).

C. DEEP SEMANTIC ATTRIBUTES

Based on the above methods, we proposed the Automated Error Detection (AED) system, as shown in Algorithm 1. The number of misspelling and grammatical errors in an essay are adopted as deep semantic attributes. The evaluation methods and results of the AGCE system will be described in the following section.

TABLE 5. The corresponding physical representation of Kappa.

Kappa	Physical presentation
0.0–0.20	almost inconsistent
0.21–0.40	generally consistent
0.41–0.60	probably consistent
0.61–0.80	highly consistent
0.81–1.0	completely consistent

VI. IMPLEMENTATION AND EVALUATION

A. DATASETS

Since Chinese essay datasets with high quality are not available, we created a standard dataset. Essays are provided by a Chinese primary school. The dataset contains around 1000 pictures of Grade 3 students' essays and each essay was pre-scored by one human expert grader. The OCR technology provided by NetEase was adopted to identify words in the picture, and the results were cleaned to preserve pinyin and semantic errors. Although there are some confusing Chinese characters in the results, we would correct them manually. In the dataset, 170 essays were randomly selected as the training set and 30 ones as the test set. The dataset is used to extract shallow and deep semantic attributes.

The Bi-LSTM-CRF model was performed on the dataset provided within the Chinese Grammar Error Diagnosis (CGED) competition. The dataset contains a total of 20451 Chinese sentences and has provided the tagging sequence including the part-of-speech tagging sequence and the grammatical error tagging sequence.

B. EVALUATION

The criteria for evaluating the AGCE system are Kappa, Linear weighted kappa (Lwk) and Quadratic weighted kappa (Qwk). These criteria are error metric that measures the degree of agreement between two graders. This approach is an analogy to the correlation coefficient. The Kappa can be calculated by

$$k = \frac{p_o - p_e}{1 - p_e}, \tag{36}$$

where p_o is the actual consistency rate, p_e is the theoretical consistency rate. The p_o could be calculated by making the number of correct classifications for each class be divided by the total number. The Kappa is usually between 0 and 1, and the corresponding physical representation is presented in Table 5.

When the difference between the prediction and the real category is greater, the consistency will be worse. For this reason, linear weighted Kappa and quadratic weighed Kappa give a solution for this problem

$$k_w = \frac{\sum_{i,j} w_{i,j} p_o - w_{i,j} p_e}{n - \sum_{i,j} w_{i,j} p_e}, \tag{37}$$

where $w_{i,j}$ is weight. If the number of essays in datasets is N and the difference of the categories is $i - j$, the linear weight

TABLE 6. Hierarchical criteria of essay scores.

Score	<60	60–67	68–77	78–87	88–93	>93
Level	1	2	3	4	5	6

TABLE 7. The training parameters of Bi-LSTM-CRF model.

Name	Parameter
batch size	64
epoch	200
embedding	100
RNN hidden dim	200
LSTM maxlen	300
dropout	0.25

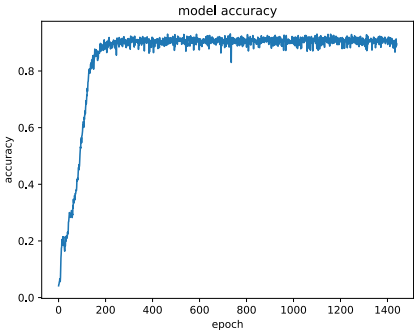


FIGURE 10. The training results of Bi-LSTM-CRF model.

and the quadratic weight can be calculated by

$$linear_w = 1 - \frac{|i - j|}{N - 1}, \tag{38}$$

$$quadratic_w = 1 - (\frac{i - j}{N - 1})^2. \tag{39}$$

In order to reasonably calculate Kappa, Lwk and Qwk, the scores predicted by the AGCE system are classified into six levels as the ASAP competition does. The hierarchical criteria are shown in Table 6.

VII. RESULTS

A. GRAMMATICAL ERROR RECOGNITION

We performed the Bi-LSTM-CRF model on the datasets provided by CGED competition. The training parameters are presented in Table 7.

The accuracy of the Bi-LSTM-CRF model on the training set is 0.9063, while on the test set is 0.9025. The sentences provided by CGED competition are complex. The inputs of our AGCE system are primary school essays, which are relatively simple. In this case, the resulting error is within the acceptable range, and we will verify this conclusion in Section VII. The training result is presented in Fig. 10.

B. EVALUATION ACCURACY OF OUR AGCE SYSTEM

We extracted and combined shallow and deep semantic attributes of 170 essays in training datasets. Then we employed the decision tree, random forest, gradient boosting decision tree and neural network to respectively fit these attributes. The test results are illustrated in Fig. 12.

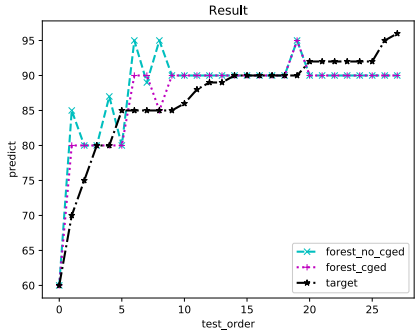
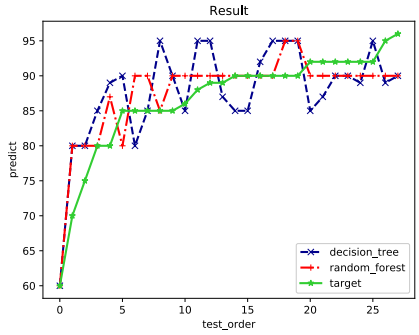
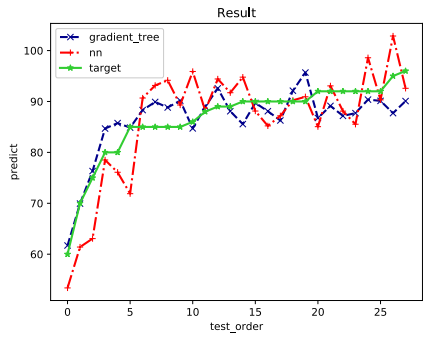


FIGURE 11. The test results of grammatical error attribute. The color of curve with adding a grammatical error attribute is maroon. The color of curve without adding grammatical error attribute is coral. The color of target curve is black.



(a) The test results of decision tree and random forest. The colors of decision tree curve, random forest curve and target curve are blue, red and green, respectively.



(b) The test results of gradient boosting decision tree and neural network. The colors of gradient boosting decision tree curve, neural network curve and target curve are blue, red and green, respectively.

FIGURE 12. The test results of AGCE system.

The errors of random forest, neural network, and gradient boosting decision tree are all less than 5 points. The prediction curve is consistent with the target curve. However, the fitting result of the decision tree is poor and its error fluctuation is large.

We test the accuracy on the test set when adding grammatical error attribute. The fitting method is random forest and the test result is illustrated in Fig. 11. This result indicates when adding a grammatical error attribute, the fitting result is better. Compared with some essays without adding the grammatical error attribute, the prediction error is reduced by about 5 points. The result also indicates the grammatical attribute extracted by the Bi-LSTM-CRF model can highly improve the accuracy of our AGCE system.

TABLE 8. The test results of our AGCE system.

Method name	Kappa	Lwk	Qwk
decision tree	0.0256	0.2343	0.4760
random forest with grammar (AGCE)	**0.4238**	**0.5845**	**0.7590**
neural network	0.2486	0.4776	0.7017
gradient boosting decision tree	0.3100	0.4567	0.6263
random forest without grammar	0.3289	0.5172	0.7210
LSTM	-0.0297	-0.0656	-0.1190

TABLE 9. The results of average error.

Method name	Average error
decision tree	5.10
random forest with grammar (AGCE)	**2.78**
neural network	3.89
gradient boosting decision tree	3.20
random forest without grammar	3.32
LSTM	10.32

TABLE 10. The comparison results between AGCE system and other popular neural networks.

Name	Qwk
AGCE	**0.7590**
LSTM	**0.756**
LSTM+attention	0.731
CNN+LSTM	0.708
BLSTM	0.699

We employ the evaluation method in Section VI to evaluate our AGCE system and the results are presented in Table 8. The evaluation values of random forest are 0.7590 (Qwk), 0.5845 (Lwk) and 0.4238 (Kappa), which is the best result in these methods. However, the evaluation values of LSTM are all negative, which indicates the deep neural network highly relies on the large corpus to improve its accuracy.

The average error between the predicted scores and the target scores is presented in Table 9. The best result comes from the random forest, which is just 2.78 points. This result can be accepted for students' use. However, the result of LSTM is up to 10.32, which is far beyond the acceptable error range.

C. COMPARISON WITH POPULAR MODELS

We compare the result of our AGCE system with other popular models, which include LSTM, LSTM+attention, CNN+LSTM, and BLSTM [36]. These models are trained in datasets provided by ASAP competition, and the comparison results are presented in Table 10. This comparison results indicate that AGCE achieves higher grading accuracy on a small dataset compared with four popular neural networks trained on large-scale datasets.

```
Keywords:[ ' 同学 ' , ' 出门 ' , ' 蹦床 ' , ' 碰碰车 ' , ' 秋游 ' ]
Length of sentences:21.0
Number of sentences:8
Grade-one vacabulary:0
Grade-two vacabulary:1
Grade-three vacabulary:0
Grade-four vacabulary:0
Grade-five vacabulary:1
Grade-six vacabulary:0
Number of words:168
Number of metaphors:0
Spelling mistakes:[[ ' spelling error: 慢天 ' , ' corrected words: 漫天 ' , '
begin: 154 ' , ' end: 156 ' ]]
Phonetic:[ " pinyin:[ ' xi ' , ' su ' ] Chinese characters:[ ' 习俗 ' ] " , "
pinyin:[ ' zao ' , ' can ' ] Chinese characters:[ ' 早餐 ' ] " ]
Score: 80
```

FIGURE 13. The feedback of an essay provided by AGCE system.

D. AN EXAMPLE OF THE PROVIDED FEEDBACK

Fig. 13 displays a simple example, which is an essay randomly selected in datasets. Our AGCE system first provides the keywords and some shallow attributes base on natural language processing tools. Then the system detects one spelling error, two pinyin and converts them into Chinese characters. Finally, the score of the essay is 80.

VIII. DISCUSSION AND CONCLUSION

In this paper, we proposed an automated Chinese essay evaluation system, AGCE, which combines the shallow and deep semantic attributes of an essay. We adopt the three most effective attributes from the IntelliMetric system, which include the number of words and sentences, the length of sentences. Regarding the Chinese text, we add the number of metaphorical sentences, pinyin and different levels of vocabulary to the shallow attributes. We compare our AGCE system with other popular neural network models. The experiment result indicates that our AGCE system achieves higher grading accuracy on a small dataset compared with four popular neural networks trained on large-scale datasets. The results of this research support the idea that our AGCE system effectively improves the utilization of samples. In addition, our AGCE system provides spelling error feedback and grammatical error feedback to students, and we find these deep semantic attributes can highly improve the accuracy of scoring. Also, we created a Chinese essay dataset, which is critical to promote the research on Chinese AEE systems.

Several questions still remain to be further investigated. The open challenges for our future work include further research of semantic attributes and constructive feedback. We plan to combine the Bi-LSTM-CRF model with an expert corpus to correct grammatical errors and improve the accuracy of the semantic error recognition. Another future challenge is to incorporate new methods for unsupervised text learning, since there are few large-scale datasets of Chinese essays.

As Chinese language tests become popular, the development of automated Chinese essay scoring systems is of great importance to teachers and students. Our AGCE system proposed in this paper helps alleviate the teachers' load, and students can use our AGCE system in the classroom as well as at home to obtain feedback about essays writing in time. Unlike some commercialized systems, we publicly provide the technical details and results of the AGCE system. It would be our honor to promote the openness of this research field and hopefully bring more Chinese AEE systems to practical applications.

ACKNOWLEDGMENT

The authors would like to present appreciation to the GuoFangKeDa Affiliated Primary School for providing students' essays to build the Chinese essay dataset. The created Chinese essay dataset and Chinese corpora are presented at https://github.com/yangyiqin-tsinghua/Automated-Grader-for-Chinese-Essay.git.

The Chinese Grammar Error Diagnosis corpora come from http://cged.biz/.

The spelling error and correction corpora come from https://github.com/shibing624/pycorrector.

REFERENCES

[1] E. B. Page, "The imminence of. . . Grading essays by computer," *Phi Delta Kappan*, vol. 47, no. 5, pp. 238–243, 1966.
[2] S. Valenti, F. Neri, and A. Cucchiarelli, "An overview of current research on automated essay grading," *J. Inf. Technol. Educ., Res.*, vol. 2, no. 1, pp. 319–330, 2003.
[3] Y. Attali and J. Burstein, "Automated essay scoring with E-rater V. 2," *J. Technol., Learn. Assessment*, vol. 4, no. 3, pp. 1–30, 2006.
[4] L. M. Rudner, V. Garcia, and C. Welch, "An evaluation of IntelliMetric essay scoring system," *J. Technol., Learn. Assessment*, vol. 4, no. 4, pp. 1–21, 2006.
[5] M. D. Shermis, C. M. Koch, and E. B. Page, "Trait ratings for automated essay grading," *Educ. Psychol. Meas.*, vol. 62, no. 1, pp. 5–18, 2002.
[6] M. D. Shermis, H. R. Mzumara, and J. Olson, "On-line grading of student essays: PEG Goes on the World Wide Web," *Assessment Eval. Higher Educ.*, vol. 26, no. 3, pp. 247–259, 2001.
[7] L. Rudner and P. Gagne, "An overview of three approaches to scoring written essays by computer," ERIC Digest number ED 458 290, 2001.
[8] J. Burstein, "The e-rater scoring engine: Automated essay scoring with natural language processing," in *Automated Essay Scoring: A Cross-Disciplinary Perspective*, M. D. Shermis, and J. Burstein, Eds. Hillsdale, NJ, USA: Lawrence Erlbaum, 2003, pp. 113–122.
[9] B. Hu, "The challenges of Chinese: A preliminary study of UK learners,' perceptions of difficulty," *Lang. Learn. J.*, vol. 38, no. 1, pp. 99–118, 2010.
[10] S. Dikli, "An overview of automated scoring of essays," *J. Technol. Learn. Assessment*, vol. 5, no. 1, pp. 1–35, 2006.
[11] P. W. Foltz, "Latent semantic analysis for text-based research," *Behav. Res. Methods, Instrum., Comput.*, vol. 28, no. 2, pp. 197–202, 1996.
[12] B. Lemaire and P. Dessus, "A system to assess the semantic content of student essays," *J. Educ. Comput. Res.*, vol. 24, no. 3, pp. 305–320, 2001.
[13] M. A. Hearst, "The debate on automated essay grading," *IEEE Intell. Syst. Appl.*, vol. 15, no. 5, pp. 22–37, Sep. 2000.
[14] J. Burstein and D. Marcu, "Benefits of modularity in an automated essay scoring system," in *Proc. COLING Workshop Using Toolsets Architectures to Build NLP Syst.*, 2000, pp. 44–50.
[15] S. Elliott, "IntelliMetric: From here to validity," in *Automated Essay Scoring: A Cross-Disciplinary Perspective*, M. D. Shermis and J. Burstein, Eds. Hillsdale, NJ, USA: Erlbaum, 2003, pp. 71–86.
[16] Z. Li, S. Link, and H. Ma, "The role of automated writing evaluation holistic scores in the ESL classroom," *System*, vol. 44, pp. 66–78, 2014, doi: 10.1016/j.system.2014.02.007.
[17] S. Elliott, "Overview of intelliMetric," in *Automated Essay Scoring: A Cross-Disciplinary Perspective*, M. D. Shermis and J. Burstein, Eds. Hillsdale, NJ, USA: Erlbaum, 2003, pp. 67–70.
[18] L. M. Ruder and T. Liang, "Automated essay scoring using Bayes,' theorem," *J. Technol., Learn. Assessment*, vol. 1, no. 2, pp. 3–21, 2002.
[19] M. T. Schultz, "The IntelliMetrix automated essay scoring engine—A review and an application to Chinese essay scoring," in *Handbook of Automated Essay Scoring: Current Applications and Future Directions*, M. D. Shermis and J. Burstein, Eds. New York, NY, USA: Routledge, 2013, pp. 89–98.

[20] I. Persing and V. Ng, "Modeling argument strength in student essays," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process.*, vol. 1, 2015, pp. 543–552.
[21] I. Persing and V. Ng, "Modeling thesis clarity in student essays," in *Proc. 51st Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2013, pp. 260–269.
[22] I. Persing and V. Ng, "Why can't you convince me? Modeling weaknesses in unpersuasive arguments," in *Proc. IJCAI*, 2017, pp. 4082–4088.
[23] I. Persing, A. Davis, and V. Ng, "Modeling organization in student essays," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2010, pp. 229–239.
[24] F. Gutierrez, D. Dou, and S. Fickas, "Providing grades and feedback for student summaries by ontology-based information extraction," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, 2012, pp. 1722–1726.
[25] X. Peng, D. Ke, and Z. Chen, "Automated Chinese essay scoring using vector space models," in *Proc. 4th Int. Univ. Commun. Symp.*, 2010, pp. 149–153.
[26] S. Hao, Y. Xu, and D. Ke, "SCESS: A WFSA-based automated simplified Chinese essay scoring system with incremental latent semantic analysis," *Natural Lang. Eng.*, vol. 22, no. 2, pp. 291–319, 2016.
[27] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes," *J. Mol. Biol.*, vol. 305, no. 3, pp. 567–580, 2001.
[28] C. H. Lee, F. Gutierrez, and D. Dou, "Calculating feature weights in Naive Bayes with Kullback-Leibler measure," in *Proc. IEEE 11th Int. Conf. Data Mining*, Dec. 2011, pp. 1146–1151.
[29] P. Rozin, S. Poritsky, and R. Sotsky, "American children with reading problems can easily learn to read English represented by Chinese characters," *Science*, vol. 171, no. 3977, pp. 1264–1267, 1971.
[30] I. I. Bejar, "A validity-based approach to quality control and assurance of automated scoring," *Assessment Educ., Princ., Policy Pract.*, vol. 18, no. 3, pp. 319–341, 2011.
[31] D. M. Williamson, X. Xi, and F. J. Breyer, "A framework for evaluation and use of automated scoring," *Educ. Meas., Issues Pract.*, vol. 31, no. 1, pp. 2–13, 2012.
[32] C. Y. Lin and E. Hovy, "Automatic evaluation of summaries using N-Gram co-occurrence statistics," in *Proc. Hum. Lang. Technol. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2003, pp. 150–157.
[33] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *J. Mach. Learn. Res.*, vol. 3, pp. 1137–1155, Feb. 2003.
[34] T. Mikolov, K. Chen, and G. Corrado, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*. [Online]. Available: https://arxiv.org/abs/1301.3781
[35] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[36] K. Taghipour and H. T. Ng, "A neural approach to automated essay scoring," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2016, pp. 1882–1891.

YIQIN YANG received the bachelor's degree in automation from Shandong University, Weihai, China, in 2019. He is currently pursuing the Ph.D. degree in system engineering with the Center for Intelligent and Networked Systems, Department of Automation, Tsinghua University, Beijing, China. His current research interests include natural language processing and reinforcement learning.

LI XIA (S'02–M'12–SM'16) received the bachelor's degree and the Ph.D. degree in control theory from Tsinghua University, Beijing, China, in 2002 and 2007, respectively. After Ph.D. graduation, he worked at the IBM Research, China, as a Research Staff Member, from 2007 to 2009, and at the King Abdullah University of Science and Technology (KAUST), Saudi Arabia, as a Postdoctoral Research Fellow, from 2009 to 2011. Then, he returned to Tsinghua University as a Lecturer, in 2011, and was promoted as an Associate Professor, in 2013. In 2019, he joined Sun Yat-Sen University as a Full Professor. He was a Visiting Scholar with Stanford University and The Hong Kong University of Science and Technology. He is currently a Professor with the Business School, Sun Yat-Sen University, Guangzhou, China. His research interests include the methodology research in stochastic learning and optimization, Markov decision processes, reinforcement learning, queueing theory, and the application research in energy systems, smart building, and financial technology. He serves as an Associate Editor of the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING, *Discrete Event Dynamic Systems*, and *Energy Informatics*.

QIANCHUAN ZHAO (M'06–SM'08) received the B.E. degree in automatic control and the B.S. degree in applied mathematics from Tsinghua University, Beijing, China, in 1992, and the M.S. and Ph.D. degrees in control theory and its applications from Tsinghua University, in 1996. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA, in 2000, and with Harvard University, Cambridge, MA, USA, in 2002. He was a Visiting Professor with Cornell University, Ithaca, NY, USA, in 2006. He is currently a Professor and the Director of the Center for Intelligent and Networked Systems, Department of Automation, Tsinghua University. He has published more than 80 research articles in peer-reviewed journals and conferences. His current research interests include the control and optimization of complex networked systems with applications in smart buildings, smart grid, and manufacturing automation. He received the 2009 China National Nature Science Award for the Project Optimization Theory and Optimization for Discrete Event Dynamic System. He serves as the Chair of the Technical Committee on Smart Buildings of the IEEE Robotics and Automation Society. He is an Associate Editor of *Journal of Optimization Theory and Applications*, the IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS, and the IEEE TRANSACTIONS ON AUTOMATION SCIENCE AND ENGINEERING.