Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 154 (2019) 13–19

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Congress of Information and Communication Technology, ICICT 2019

Research on Text Error Detection and Repair Method Based on Online Learning Community

Xu Song[a,*], Ye Jun Min[a], Luo Da-Xiong[a], Wang Zhi Feng[b], Chen Shu[a]

[a]School of Computer, Central China Normal University, Hubei Wuhan 430070, China
[b]School of Educational Information Technology, Central China Normal University, Hubei Wuhan 430070, China

Abstract

The short text in the online learning community is an important source of data in learning analysis. Therefore, the quality of the short text has a significant impact on the study of learning analysis. Due to the large amount of text data in the learning community, manual detection and repair will cost too much. This paper proposes a text detection and repair framework based on an online learning community. It aims to automatically detect and repair various types of semantic errors and grammatical errors that exist in online learning community short texts. The framework utilizes existing text error detection and repair algorithms and integrates them effectively to form a comprehensive detection and repair algorithm. In this paper, the validity of the framework is verified through experiments on the constructed data set. The experimental results show that the framework has high accuracy in automatically detecting and repairing text errors.

Keywords: Online learning community; Chinese text errors; Semantic error detection; Grammatical error detection; Text error correction

1. Introduction

The online learning community refers to the current popular online learning platform. The daily generated learning data is very large. Learning and analyzing these data can improve the personalized learning support services

* Corresponding author. Tel.: 15623010207.
E-mail address: 798815708@qq.com

1877-0509
10.1016/j.procs.2019.06.004

of the learning platform and help users improve learning efficiency. Most of these data are short-text data (such as user questions, answers, comments, etc.). Therefore, in order to ensure the validity of short-text based learning analysis research, the study of textual error detection and repair of short texts in online learning communities is very It is necessary[1].

Text errors mainly include semantic errors (such as real-word errors) and grammatical structure errors[2]. These errors are usually caused by user's input errors or insufficient user knowledge. Detecting and fixing these errors will help improve the effectiveness of learning analysis and research and ensure the successful application of learning and analysis techniques.

2. Text Error Detection and Repair Framework

Currently, there are many methods for detecting text errors. There are n-gram-based probabilistic statistical methods[3], methods based on contextual context[4], and long-distance word-based fixed collocations for detecting and fixing real-word errors[5] and so on. The method for the grammatical structure of the wrong method based on natural language processing techniques[6]. These methods all have their own advantages and disadvantages. Therefore, how to make full use of the advantages of these methods for the detection and repair of text errors is a problem worthy of further study. In order to improve the accuracy of text error detection and repair and to detect and repair more different types of text errors, this paper integrates existing detection and repair methods and proposes a text error detection and repair framework, as shown in Fig 1.

This framework is mainly to integrate two types of error detection and repair methods, namely semantic error and syntax error detection and repair methods. The first step is to pre-process, segment the text and construct the confusion set based on the text corpus; then, according to the confusion set, determine the homonym in the sentence. If it is a homonym, use the n-gram probability statistics method, contextual context statistics method, and word collocation. Methods are used to make true word misjudgment. At the same time, the use of part-of-speech rules method is used to detect whether there is a grammatical structure error, and the detected grammatical structural errors are marked and repaired. The framework uses the text to be detected as input data and the detected and repaired text as output data.

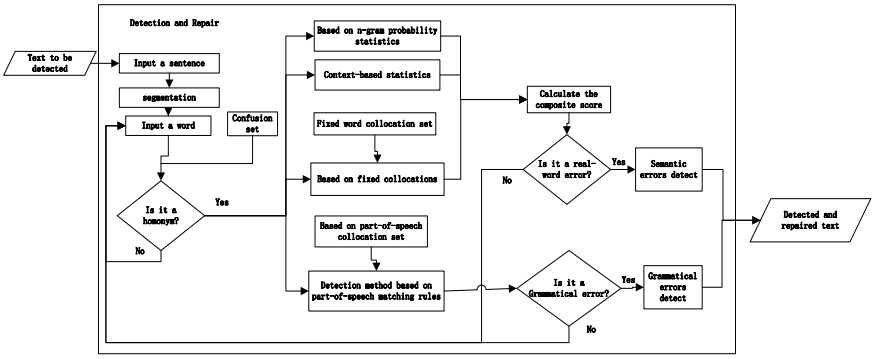


Figure. 1. Text error detection and repair framework.

3. Text Error Detection and Repair Method Design

3.1. Semantic error detection and repair method

The semantic errors in the online learning community mainly refer to real-word errors, which means miswriting a word as another valid word in the dictionary.

The algorithm idea based on n-gram probabilistic statistical model is to count the sequence's frequency of homonyms and n words before and after in the text corpora. Then replace the homonyms in the sentences with the

words in the confusion, count the sequence's frequency of each confusion word and n words before and after in the text corpora[7]. Then the maximum likelihood estimation is used to calculate the n-gram probability[8]. The n-gram method generally considers a low-probability sequence to be erroneous, and a high-probability sequence is a candidate list for error correction suggestions.

The algorithm concept based on the contextual context model is to calculate the sum of frequencies in the text corpora by counting the 2-gram model consisting of the homonym and the adjacent top k words and the following k words in the sentence. Then determine whether the homonym is real-word error based on the frequency level[2].

The idea of the Chinese fixed collocation algorithm is to automatically construct a collocation knowledge base[5] to process every sentence in the text of the waiting pair, find the homonym existing in the sentence, find the collocation information of the homonym, calculate the degree of collocation support of the homonym and each confusion word.

The above three algorithms are designed with different influencing factors and measurement standards. In terms of influencing factors, n-gram's probabilistic statistics mainly regards the frequency of the sequence of the original word and the surrounding words as the influencing factor, while the context of the context regards the frequency of the co-occurrence of the original word and the surrounding word as the influencing factor. Chinese fixed collocation uses the collocation frequency between the original word and the long-distance word as the influencing factor. In the measurement standard, the value range is in the range of [0,1]. The contextual context and Chinese fixed collocations are based on the frequency size as a standard, and the value range is an integer greater than or equal to one. To unify the metrics of the above algorithm results, the above calculated scores can be recorded as S1, S2, and S3. In this paper, based on the experimental and empirical values, the relevant weights are set to a1=0.4, a2=0.3, a3=0.3, so that the scores can be weighted and summed using formula (1) to obtain a comprehensive score for each homonym and confusion word. Calculate the result S according to formula (1) as a measure of whether the homonym W_i is a real-word error.

$$S(w_i) = a1 * S1 + a2 * S2 + a3 * S3 \quad (1)$$

3.2. Grammatical errors detection and repair methods

1) Constructing Part-of-Speech Binary Group

The relative position of each word in a sentence is closely related to the part-of-speech of the preceding word and the word of the latter word[5]. For this purpose, each sentence in the data set is pre-processed (ie, word segmentation and POS tagging), and then a word is selected as the center word Wi, and the previous word adjacent to the word is selected to form a part-of-speech binary group (Wi-1, Wi), or select the next word adjacent to the word to form a part-of-speech binary group (Wi, Wi+1). If the center word is the first word, the previous word is set as the character "#begin#"; if the center word is the last word, the next word is set as the character "#end#".

2) Constructing a Set of Part-of-Speech Matching Rules

According to the linguistic center word analysis method[9], when two or more words are collocated, there must be one word as the center word and the other word as the collocation word. To this end, we will use nouns, verbs, adjectives and adverbs as the central words to construct a set of parts of speech rules R. Based on the online learning community, this paper constructs 12 rules of part-of-speech matching, as shown in Table I.

Table 1. Parts-of-speech matching rule set.

No	Rules	Example
1	N + [N, NR, NS, NT, NZ]	Dad + Mam
2	N + [V, VD, VN]	teacher + teach
3	N + [A, AD, AN]	landscape + beautiful
4	V + [N, NR, NS, NT, NZ]	listen + music
5	V + [V, VD, VN]	sing + dance
6	V + [D, DG]	run + quickly
7	V + Q	have + a
8	A + [N, NR, NS, NT, NZ]	beautiful + flower
9	A + [A, AD, AN]	smart + clever
10	Q + [N, NR, NS, NT]	an + apple
11	D + [V, VD, VN]	quickly + eat
12	D + [A, AD, AN]	quite + angry

The contents in square brackets are optional contents, and the meanings of related symbols are ordinary nouns (remembered as N); names of people (remembered as NR); place names (remembered as NS); institutional groups (remembered as NT); other proper names (remembered as NZ).

3) Constructing Collocation Knowledge Base

The structure of collocation knowledge base should satisfy three requirements[10]: (1) the extracted collocations should be as accurate as possible; (2) the extracted collocations can be measured using collocation strengths (such as mutual information, etc.); (3) These collocations can be categorized in language to facilitate subsequent collocation corrections.

Algorithm 1. Detection and repair for grammar structure error based on part-of-speech.

Algorithm function: Detect grammatical errors in short text, mark and repair.

Input: To be detected and repaired text T1; Rules Set R; $\alpha 1, \alpha 2$ (the weight of MI and PD).

Output: Detected and repaired Text T2.

Begin

1. segmentation; $ST = W_1 W_2 \ldots W_i \ldots W_n$;
2. While $0 < i < len(ST)$ /* $len(ST)$ is the length of ST*/
3. Extract the binary group of part-of-speech about W_i ;
4. if W_i is noun (or verb, or adjective) (set collocation as $c = n1 + n2, n1$ is center word)
5. Ergodic collocation rules in R about noun (or verb, or adjective) ;
6. if C in R
7. C is true collocation;
8. else
9. C is grammar structure error;
10. Find candidate collocation list1 about n1 in collocation knowledge base, calculate MI and PD;
11. calculate $S = \alpha 1 * MI + \alpha 2 * PD$;
12. replace c with C_0 which is highest score.
13. i++;
14. output: T2.

End.

Figure.2. Detection and repair algorithm for grammar structure error based on part-of-speech

According to the part-of-speech matching rule set R, word collocation extraction is performed on the short texts in the online learning community. The n-gram probabilistic statistical model is used to perform probability statistics on collocations in sentences. When the probability value is greater than the threshold k (according to the corpus level adjustment), the word collocations are stored in the collocation knowledge base under the corresponding part-of-speech matching rules.

4) Detection and Repair of Grammatical Structure Error Based on Parts of Speech Collocation

Based on the collocation knowledge base and rule set, this paper proposes a grammatical structure error detection and repair algorithm based on part-of-speech matching. The idea of this method is divided into two steps: (1) Detection of grammatical errors. Segmentation of short texts and part-of-speech tagging; word-by-word scanning, extraction of part-of-speech binaries, use rules in rule set R to determine grammatical structures to detect grammatical structural errors, and (2) make detected grammatical errors repairs. If the collocation detection is grammatical structure error, then the candidate collocation list List1 corresponding to each word is obtained in the collocation knowledge base, and then the idea of maximum likelihood estimation is used to calculate the collocation mutual information MI and collocation polymerization degree PD[5]. And according to formula (2) calculates the mixed value S of these two values, through the experiment analysis sets the parameter to 0.3 and 0.7 respectively, and sorts the collocations in the candidate collocation list according to the mixed value S. The specific algorithm is shown in Fig 2, where the sentence length is N and the collocation knowledge base size is M. The time complexity of the algorithm is O(N*M).

$$S = \alpha 1 * MI + \alpha 2 * PD \tag{2}$$

4. Detection and Repair Effect Evaluation

4.1. Experimental preparation

The data used in this experiment comes from a machine learning section in the community called Zhihu. We obtained a total of 2653 sentences and obtained 1500 comparisons after preprocessing (removing pictures, symbols, and short sentences such as "good" and "very good"). We manually construct 500 semantically incorrect sentences and 500 sentences with incorrect grammatical structure on this data. Then we use the constructed data set as the experimental test set. The research in this paper aims at the detection and repair of text errors. Therefore, the experiment is divided into two steps: the first step is to detect and analyze the text errors; the second step is to repair and analyze the detected text errors. In order to compare the effect of the size of the data set on the experimental results, experiments were conducted using data sets of 500 sentences, 1000 sentences and 1500 sentences in three different scales.

The evaluation indicators used in the experimental analysis include RECALL, ACCURACY, and F1, where RECALL is the ratio of the number of errors that are correctly detected (repaired) to the number of actual errors, and ACCURACY is the number of errors and detections that are correctly detected (fixed). Fix) the ratio of the number of errors. The calculation formula for each assessment index is as follows:

$$\text{RECALL：R}=\frac{\text{The number of correctly detected(repaired) errors}}{\text{Actual number of errors}}\times 100\% \tag{3}$$

$$\text{ACCURACY：P}=\frac{\text{The number of correctly detected(repaired) errors}}{\text{The number of detected(repaired) errors}}\times 100\% \tag{4}$$

$$\text{F1：C}=\frac{2*P*R}{P+R}\times 100\% \tag{5}$$

Table 2. Semantic error detection.

	1	2	3
RECALL	76.9%	78.7%	79.6%
ACCURACY	86.0%	87.2%	90.2%
F1	81.2%	82.7%	84.6%

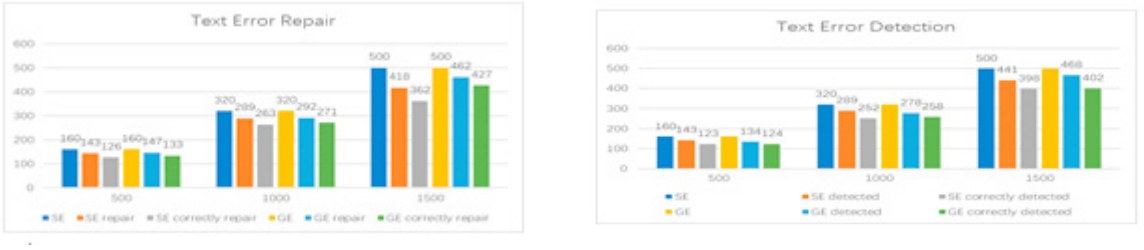
Table 3. Grammatical error detection.

	1	2	3
RECALL	77.5%	80.6%	80.4%
ACCURACY	92.5%	92.8%	93.6%
F1	84.3%	86.3%	86.5%

4.2. Text error detection

In the first test experiment, the error detection *RECALL*, *ACCURACY*, and *F1* result in the experiment are shown in Table 2, 3 and fig.3(a), semantic error is denoted as SE, grammar error is marked as GE:

It can be seen from Tables 2 and 3 that the framework proposed in this paper is effective for the detection of semantic errors and grammatical errors. The accuracy of detection is about 90%. And with the increase of the data set, the accuracy has also gradually improved. This is because the detection method is based on the probability of co-occurrence of words or parts of speech in the corpus.



(a) (b)

Fig.3 (a)Text Error Detection;(b) Text Error Repair.

4.3. Text Error repair

In the second step of the repair experiment, the recall rate, accuracy rate and F1 result of the error recovery in the experiment are shown in Tables 4 ,5 and fig.3(b).

Table 4. Semantic error repair.

	1	2	3
RECALL	78.7%	82.1%	83.6%
ACCURACY	88.1%	91.0%	92.6%
F1	83.1%	86.3%	87.9%

Table 5. Grammatical error repair.

	1	2	3
RECALL	83.1%	84.6%	85.4%
ACCURACY	90.4%	92.8%	92.4%
F1	86.6%	88.5%	88.8%

It can be seen from Tables 4 and 5 that the framework proposed in this paper is effective for the repair of semantic errors and grammatical errors. And with the increase of data sets, the accuracy of the repair is gradually improved. This is because the detection is based on the co-occurrence probability of words or parts of speech in the corpus, and the repair of this article is based on the results of the test.

5. Conclusion

This paper proposes a Chinese text error detection and repair framework based on an online learning community. It aims to solve the problem of detecting and fixing text errors in the online learning community. It automatically detects and repairs semantic errors and grammatical errors in short texts, reduces the cost of manual detection and repair, and can also be based on online learning communities. Analytical research provides a better source of data. The construction of this framework helps to combine the advantages of existing methods, improve the accuracy of text error detection and repair, and the number of types of errors. According to the experimental results designed in this paper, the framework has high accuracy in the detection and repair of text errors and can effectively detect and repair semantic errors and grammatical errors.

At present, the work of this paper only solves the detection and repair of semantic and grammatical errors. In the future, it can also expand the detection and repair methods of other types of errors, making the detection and repair of textual errors more comprehensive and accurate.

Acknowledgement

This work was supported in part by the National Social Science Fund General Project (17BTQ061).

References

1. S. Sharma, S. Gupta. A Correction Model for Real-word Errors[J]. Procedia Computer Science, 2015, 70: 99-106.
2. C. Li, H. Liu. Association analysis and N-Gram model based approach to detecting incorrect arguments[J]. Ruan Jian Xue Bao/Journal of Software, 2018,29(8) (in Chinese). http://www.jos.org.cn/1000-9825/5531.htm.
3. L. L. Liu, S. Wang and D. S. Wang.Automatic Text Error Detection in Domain Question Answering[J]. Journal of Chinese Information Processing, 2013, 27 (3):77-83.
4. L. Wang, Y. S. Zhang, Quantitative Analysis and Extracting Arithmetic 39(Z6):232-234,270.
5. M. Yu, T. S. Yao. A Hybrid Method for Chinese Text Collocation[J]. Journal of Chinese Information Processing, 1998, 12 (2):31-36.
6. Z. Xie, A. Avati and N. Arivazhagan, et al. Neural language correction with character-based attention. 2016, arXiv preprint arXiv:1603.09727.
7. M. S. Dashti. Real-word error correction with trigrams: correcting multiple errors in a sentence[J]. Language Resources & Evaluation, 2017(2):1-18.
8. L. L. Liu, C. G. Cao. Chinese Real-word Error Automatic Proofreading Based on Combining of Local Context Features[J]. Computer Science, 2016,43(12):30-35.
9. Y. S. Zhang, J. Zhang. Study of Semantic Error Detection Method for Chinese Text[J]. Chinses Journal of Computers, 2017, 40(4):911-924.
10. C. R. Li. Research and Design of a Kind of Hierarchical Language Model for English Grammar Error Correction[D]. University of Science and Technology of China, 2017.