Detection of Non-Standard English Expressions by Language Sense

Li Chao and Zhang Jingxiang*
College of Information Science and Engineering, University of Jinan, Jinan 250022, Shandong Province, China

*Corresponding author’s e-mail: ise_zhangjx@ujn.edu.cn

Abstract. This paper proposes an algorithm for perception English essay, in order to improve the ability of Chinese students’ writing which expressing more standard. Because students are influenced by their mother tongue in second language acquisition, English expression is not accurate enough. From the lexical and syntactic aspects, this paper uses the language sense relevance analysis algorithm (N-LanSen) based on the N-model to deeply perceive the students’ inappropriate expression in the essay. Different extraction methods are used to extract different features, and the corresponding feedback can be given to the places where the expression is not standard. The experimental results show that the algorithm can detect the place where English expression is not standard with high accuracy.

1. Introduction
Artificial intelligence processing is replacing the routine repetitive work handled by traditional manpower in many fields, at least it assisting manpower to carry out efficient work. English automatic scoring is no exception. At present, the number of students is much higher than that of teachers under all levels of education system in China. English teachers are unable to complete extensive corrections, while students have a strong demand in this respect, because it is difficult for students to complete a task by themselves through comparing standard answers. At present, there are three main stages in the foreign system of English essay scoring: firstly, PEG system can get scores indirectly by counting some indicators of the essay, but it can’t understand the content of the composition, only simulate the scores similar to the manual marking according to the text characteristics [1]; secondly, the natural language processing technology has developed relatively, and E-rater scored by examining the organizational structure, vocabulary and syntactic structure of the essay [2]; Writing Roadmap system tends to automatically score and personalized comments [3]. Foreign automatic writing scoring system is mainly designed for native speakers of English. For Chinese students, English is a non-native language and an official language. The above foreign systems can’t satisfy their assessment of English writing for the main purpose of foreign language learning.

There are also some automatic scoring systems for English essay in China. According to the principle of classification, the typical examples are the system of similarity comparison based on large English corpus to get the score of composition evaluation [4], and the system of feature composition evaluation based on words, sentences and textual structure to get the score through Adaboost algorithm [5]. However, these algorithms have not been studied from the aspect of English language sense to test whether English expression is standard or not. With more and more frequent exchanges between countries, it may affect communication with each other which English expression is not standard.
This paper studies the problems existing in English writing based on this demand. It proposes a method of N-language sense correlation analysis, which can effectively detect the problem of essays which English expressions isn’t standard. And give feedback to the existing problems in order to improve students’ English expression level.

2. Language sense detection algorithms and non-standard English perception

2.1. Language sense detection algorithms
N-gram is the existing language model, which uses conditional probability to evaluate the reasonableness of sentences [6]. The N-gram model is based on the Hidden Markov Hypothesis, which assumes that the occurrence of the nth word is only related to preceding n-1 word, but not to other words. The probability of the whole sentence is the product of the probability of each word appearing in the corpus. According to the probability statistics and calculation of sentences, if the probability is less than a certain threshold, it is considered that there are grammatical errors in sentences. In the N-gram model, N isn’t the larger the better, the larger N the more probability calculations. Although the current computer has been able to solve the calculation problems caused by the larger N value, but the larger N value will lead to data sparsity, the overall performance can’t be improved.

The N-gram model calculates the probability of words appearing in the whole sentence, and the result is only that the sentence is judged to be grammatical error because the probability is lower than the threshold value. It is also necessary to smooth the N-tuple in the corpus so that the detected sentences can’t occur at zero probability due to the absence of some phrases in the corpus.

The data in the corpus do not necessarily contain the expression of human language as a whole, and it is normal for phrases to be missing. However, the N-gram model adopts a smoothing method to deal with phrase missing, which is not necessarily reliable. For example, there are grammatical errors in sentences themselves, but the probability of sentences after smoothing is lower than the threshold of grammatical errors.

Some literatures interpret language sense as: sensitivity to following or deviating from the established usage of a language; sense of linguistic validity or appropriateness [7]. According to this idea, there should be a certain correlation between the words in the sentence. In this paper, an association rule analysis algorithm based on N-tuple, N-LanSen, is proposed for language sense measurement.

The corpus and the sentences to be tested are divided into N-tuples, and the N-tuples of the sentences to be tested are used as test sets of association rules, and the linguistic sense values of N-tuples are calculated. If the linguistic perception is less than the given threshold, the sentence expression will be problematic.

2.2. Vocabulary perception
Looking at errors in English essay, inappropriate vocabulary expression is a common problem. This is largely due to the use of English-Chinese dictionaries by students [8]. In the process of learning English, the Chinese explanations in dictionaries are regarded as the real meaning of a word, or the words are understood according to the Chinese explanations. These two approaches are unreliable. Because of the great differences between China and the West, the use of words is not the same. Using Chinese thinking to understand English vocabulary is the main reason for its mistakes. English-English dictionaries should be used as much as possible in English learning. Explaining English words in English can not only enhance the sense of English, but also realize the true meaning of English words. The perception of non-standardized vocabulary is mainly reflected in the following aspects.

2.2.1. Lexical redundant. In the process of writing, the use of words is obviously redundant, especially when they express their personal opinions. They usually use the expression “Personally speaking” or “In my opinion” with “I think”, it shows that it is redundant. Repetition is a rhetorical device in Chinese, while repetition is taboo in English [9].
Similar lexical redundancy is also reflected in the repetition of meaning caused by the use of a word and its derivatives together. The word “solve” is the verb form in English, and its derivative “solution” is a noun. “Solve a solution” is a correct expression in sentences, but “solution to solve the problem of air pollution” is redundant. When the verb "solve" is removed, it does not affect the expression of the sentence at all, and the sentence become simpler and clearer.

The use of redundant words in writing will give teachers a bad impression. First, teachers feel that students have no sense of authentic English, but they translate it directly from Chinese to English. Some teachers also feel that the redundant expressions of students may be adding the number of words, showing a phenomenon of insufficient English ability.

If the threshold of language sense is lower than the threshold of language sense, the relevant N-tuple keywords are extracted according to the rules of the same part of speech, and the cosine similarity analysis is used. If the similarity between the two N-tuples is high, the redundancy of words is considered.

Cosine similarity is the calculation of text similarity. It maps words into space by TF-IDF, and also use kick, and the correct rate is not necessarily high. Therefore, this paper adopts it. To solve this problem, this paper adopts the strategy of editing distance. The collocation of verbs and nouns in Chinese and English is not necessarily right to put it in the current semantics. Regardless of semantics, the verb collocation of feedback is definitely right. But it exist in the corpus. However, the number of words is high, the redundancy of words is considered.

Cosine similarity is the calculation of text similarity. It maps words into space by TF-IDF algorithm to obtain vector representation, and then calculates vector similarity by using cosine theorem [10]. The higher the similarity, the closer the cosine value is to 1, otherwise the cosine value is 0. The expression of cosine similarity is as follows: 

$$\cos(\theta) = \frac{a \cdot b}{\|a\| \times \|b\|}.$$  

2.2.2. Improper collocation of vocabulary. The collocation of verbs and nouns in Chinese and English is different [11]. In daily life, the words "eat" can be used in meals and medicine in Chinese expressions are no more common, and also use "kick football". But in English expression, if we use “eat meals” to express eating; “eat medicine” to express taking medicine; “kick football” to express playing football, this is the so-called "apply mechanically" and translate directly according to Chinese thought. Generally, playing is used in sports such as “kick” and “play”, while Chinese is more detailed. The English language is confused in eating. We can’t use “eat medicine” to express "take medicine", only when eating specific food, use eat, such as eat apples.

The usage of English vocabulary is quite complex. If there is no sense of language, it is easy to make mistakes just by expressing ideas in Chinese. It will be a huge task to use rules to express English in essays which are not standard, and the correct rate is not necessarily high. Therefore, this paper adopts a statistical method to detect the improper collocation of English vocabulary.

For a certain N-tuple, the value of language sense is below the threshold, and it happens that there are verbs and nouns in the N-tuple, which is suspected of improper lexical collocation. By extracting the noun in the sentence and searching the corpus for the collocation of the noun. Since there is more than one collocation of a noun, such as "play football" mentioned above, the word "buy football" may exist in the corpus. If the threshold of language sense is lower than the threshold of language sense, the verb collocation of feedback is definitely right. But it is not necessarily right to put it in the current semantics. To solve this problem, this paper adopts the strategy of editing distance. The Chinese expression of "kick football" in English is "play football"; while "buy a football" is not "buy football" in English, the article “a” must be used in the middle, otherwise it is also wrong.

Editing distance is a distance measurement method. It is the number of editing operations needed to convert one character into another character [12]. The expressions of "buy a football" and "play football" discussed in this paper need to be replaced by the verb “kick”, while "buy a football" needs to add the article “a”, so the editing distance is longer than "play football".

Of course, the use of editorial distance is only an idealized feedback without any other grammatical errors in the essay. If a student writes not "play football" but "buy a football", using editorial distance will not solve the problem. Idealized use of editing distance is not very reliable, but also need to extract the semantics of the sentences before and after, to obtain the correct vocabulary collocation.

2.3. Syntactic perception
2.3.1. Misuse of conjunctions. There are also great differences between English and Chinese in conjunction [13]. In Chinese, causality is often expressed by "because … so... ". It is easy for students with weak English sense to use “because” and “so” in the same sentence when expressing causality. This is not in accordance with grammatical rules in English expression. In a sentence, conjunctions can't exist at the same time when we use "because … So... ", similar also includes "although… but... ".

For the detection of conjunctions that can't be used together, if we use N-LanSen directly, it is very likely that it can’t be detected. Because "because" needs to be followed by a clause expressing the cause, if the clause is longer, according to the rule: the occurrence of the N word is only related to its preceding N-1 word. Then the N will be larger, the data sparse, or nonexistent. In order to avoid large N value and avoid sparse or non-existent data, this paper extracts the key words in sentences and uses the sense of speech analysis based on specific parts of speech.

By using clause operation in essay, we can get a complete sentence to be tested. To extract the conjunctive words in sentences, the N-gram model is used to extract the words. For example:

“Because it’s raining outside, so I won’t go out.”

Words with conjunctive attributes extracted from part of speech are “because” and “so”. The two words are put into the binary grammar model < because, so>. Since there are no such binary groups as < because, so > in the standard corpus, we can consider the use of conjunctions errors.

2.3.2. Error of passive voice. Active voice is often used in Chinese expression, but passive voice is seldom used. In Chinese, the passive voice will not be used if the sentence does not need to point out the executor of the action and there is no ambiguity between the active and passive expressions [14]. In terms of English expression, it is contrary to Chinese expression. In English expressive thinking, subjectivity is generally minimized, and the beginning of "person" is replaced by the beginning of "material name".

The conversion of active voice to passive voice will lead to great changes in sentence structure, and the use of active voice in some sentences is also in line with the habit of English expression. Therefore, this paper only detects the passive voice in the aspect of voice, and reflects whether the candidates are standardized in English writing through the aspect of passive voice.

According to various sentences with passive voice, we can find that the passive form in sentences is mainly "be + transitive verb past participle". In fact, the “be” here is meaningless. It only involves changes in person, number and tense.

When detecting the passive voice in a sentence, all verbs in the sentence are extracted according to the part of speech and combined into N-tuples. N-LanSen algorithm is used to test whether the sense of language in sentences is correct. For example: This book will be published.

The verbs in the sentence include “will, be, published”. Combining them into a triple < will, be, published > and matching them in the corpus. If the sense value of the triple in the corpus exceeds the minimum threshold, the sentence writing is correct.

3. Algorithmic flows and statistical analysis of experiments

3.1. N-LanSen model

This model uses a sliding window based on part-of-speech markers to extract words, and it also needs to do clause processing before detecting the essay [15]. For example, when evaluating the sense of the sentence "He is reading English." starting with the word “He”, only two words are extracted from the sentence at a time. After the extraction of the current two words “He” and “is” is completed, the extraction of the words is started iteratively. Finally, the result of binary model is < He, is > < is, reading > < reading, English >. The binary model is established as shown in Figure 1.

While extracting non-standard English syntactic features part-of-speech tags are needed to acquire N-tuples according to specific parts of speech.

For example, we use the sentences shown in Figure 1 to extract verbs. When we need to extract verbs, we ignore other parts of speech and get a binary < is, reading >. Similarly, if a noun needs to be
extracted, only the binary group < He, English > composed of the words “He” and “English” is obtained. Similarly, according to feature extraction requirements, we can extract triples and so on.

| sentence | He | is | reading | English |
|----------|----|----|---------|---------|
| POS      | Pron. | V. | V. | N. | end |
| tuple    | He | is |
| tuple    | is | reading |
| tuple    | reading | English |
| result   | <He, is> | <is, reading> | <reading, English> |

**Figure 1.** Tuples set-up model.

After extracting N-tuples from the sliding window, the language sense can be calculated according to the data in the corpus.

In this model, the words extracted from the N-tuple are related, which can be called language sense. For example, the tuple < He, is > extracted in Figure 1 shows that a person with strong language sense knows that the verb “be” uses “is” rather than “am” or “are” without thinking.

The extracted N-tuples are compared according to the corpus. If they exceed a certain threshold, the sense of language is correct; otherwise the sense of language is wrong. Taking the calculation of sense value of binary tuple as an example, the calculation formula is (1).

\[
VLanSen(2\text{tuples}) = \frac{\text{CorpusAppearTimesWith } <a, b>}{\text{CorpusAppearTimesWith } <a, x>}
\]

Where the expression <a,x> in the form begins with “a” and ends with any word.

3.2. Statistical analysis of experiments

This experiment use the sentences of essay which written by high school students as dataset. Some of these sentences have non-standard expressions, some are proper expression. In these non-standard sentences, there may be one or more of the above problems. Firstly, these data are labeled manually to record the problems in sentences to the database. After the system language sense detection, the correct rate of the test results is counted.

The results are shown in Table 1 by detecting N-LanSen of the sentences to be tested in the database.

**Table 1.** Detection result.

| feature                          | accuracy  |
|----------------------------------|-----------|
| proper sentence                  | 100%      |
| lexical redundant                | 95.17%    |
| improper collocation of vocabulary | 97.35%   |
| misuse of conjunctions           | 99.23%    |
| misuse of passive voice          | 99.85%    |

When the system detects all the appropriate sentences, the system prompts that the sentences express correctly. But for the sentences which expressions are not so standard, the accuracy of detection isn’t completely correct.

The purpose of this paper is not only to find out the shortcomings of non-standard English, but also to improve students’ standardized English expression. In order to improve the standard of English expression of Chinese students, it is suggested that students should read more authentic English articles in English learning. Learning standard English can strengthen the sense of language, rather
than reciting beautiful sentences or essays mechanically. Because the system adopts different feature extraction methods according to different types, it can feedback the corresponding modification suggestions according to different types when feeding back the composition results to students. The system feedback is shown in Table 2.

| feature                              | feedback sentence                                                                 |
|--------------------------------------|-----------------------------------------------------------------------------------|
| lexical redundant                    | Sentence(XXX), the word "a" and "b" is redundant                                  |
| improper collocation of vocabulary   | Sentence(XXX), the verb "a" and noun "b" is improper collocation                   |
| misuse of conjunctions               | Sentence(XXX), the conjunction "a" is misuse                                       |
| error of passive voice               | Sentence(XXX), the passive voice is error                                          |

4. Conclusions
At present, N-gram is the existing grammar model, because of its large granularity, it can only get the whole sentence is correct or not, and it can’t be sure that there are mistakes in some places. Starting from the smaller granularity, this paper finds out the problems existing in sentences, and uses different detection methods according to different types of errors. The results show that the improper sentence expression can be obtained by using N-LanSen algorithm, and the recognition accuracy is high. It can meet the requirement of feature extraction and feedback the inappropriate expressions in English compositions.

Reference
[1] McNamara D S, Crossley S A and Roscoe R D 2015 J. A hierarchical classification approach to automated essay scoring. Assessing Writing. 23 35-59
[2] Shermis M D, Burstein J, Higgins D and Zechner K 2010 M. Automated Essay Scoring: Writing Assessment and Instruction. International Encyclopedia of Education (Third Edition). 20-26
[3] Xiaodong Lei 2015 J. Domestic Research and Application of Automatic Assessment System for English Compositions. Technology vision. 35 43-45
[4] Xuliang He 2013 J. A study on the reliability and validity of Juku's English composition score. Modern Educational Technology. 23 64-67
[5] Ting Li and Jingxiang Zhang 2018 J. Centralized Trend Adaptive Enhanced English Composition Scoring Algorithms. Computer Engineering and Application. 54 151-5+256.
[6] Garner J, Crossley S, and Kyle K 2019 J. N-gram measures and L2 writing proficiency. System. 176-187
[7] Jingxiang Zhang 2000 J. Talking about English Language Sense. Foreign Language Teaching. 1 88-92
[8] Macaro E and Tian L 2015 J. Exploring teachers' oral explanations of new English lexical items in a Chinese university: Comparisons with dictionary information. System. 52 78-90
[9] Jianping Zeng 2018 J. Provisional Translation of Chinese Semantic Repetitive Words and Sentences. Chinese Science and Technology Translation. 31 37-40+33
[10] Al-Anzi F S and AbuZeina D 2017 J. Toward an enhanced Arabic text classification using cosine similarity and Latent Semantic Indexing. Journal of King Saud University - Computer and Information Sciences. 29 189-195
[11] Gulec N and Gulec B A 2015 J. Lexical Collocations (Verb + Noun) Across Written Academic Genres in English. Procedia - Social and Behavioral Sciences. 182 433-440
[12] Martin R R 2015 J. On the computation of edit distance functions. Discrete Mathematics. 338 291-305
[13] Martinez A C L 2015 J. Use of Conjunctions in the Compositions of Secondary Education Student. Procedia - Social and Behavioral Sciences. 212 42-46
[14] Zhi-ying LIU and Yao-hong JIN 2012 J. Passive sentence transformation in Chinese-English patent machine translation. The Journal of China Universities of Posts and Telecommunications. 19 135-9
[15] Ranucci G 2006 J. On the significance of signal search through the “sliding window” algorithm. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. 562 433-8