Enhancing Automatic Chinese Essay Scoring System from Figures-of-Speech

Tao-Hsing Chang[1,2], Chia-Hoang Lee[2], and Yu-Ming Chang[2]

[1] Research Center for Psychological and Educational Testing, National Taiwan Normal University, 10610 Taipei, Taiwan
[2] Institute of Computer Science and Engineering, National Chiao Tung University, 30010 Hsinchu, Taiwan
{thchang, chl, Porter}@cis.nctu.edu.tw

Abstract. Chinese automated essay scoring (CAES) is a very important tool for many educational researches. However, none of the methods for retrieving features in English essays is applicable to Chinese writings because Chinese grammar parsers cannot produce reliable and useful syntactic features. CAES systems must explore other distinct features in Chinese writing. This paper proposes a method for retrieving two Chinese figures-of-speech from essays and to predict the scores of essays based on the two skills used in the essays. A scoring model based on ID3 decision tree algorithm will validate and show that the proposed method increases the performance of CAES effectively.

Keywords: Automated essay scoring; Figures-of-speech; Chinese writing; ID3.

1 Introduction

Automated essay scoring (AES) in Chinese is a very important tool for such educational researches as curriculum and instruction, psychometrics, and educational testing. The studies on above domains usually use large writing samples to analyze the relationship between the score and the writer of an essay. The task, requiring many certified raters or experts, is very costly and time consuming and therefore impedes the development of many important and interesting studies on these domains. It is evident and urgent to develop theories and techniques for machine to automatically grade essays.

Many studies ([1], [2], [3], [4], and [5]) for AES in English have been proposed and successfully applied to business products because of breakthrough in the field of natural language processing and information retrieval [6]. These approaches mainly contain three phases: retrieving the features of essays, modeling the associations between the score and the features in essays, and constructing a model for predicting essay scores using machine-learning methods. However, none of the methods for retrieving features in English essays is applicable to Chinese writings. This is due to a loose and ambiguous definition of Chinese sentence. Furthermore, current Chinese grammar parsers cannot produce reliable and useful syntactic features due to its low precision rates. Hence, the technique for developing CAES systems must explore other distinct features in Chinese writings.

Many studies for Chinese writing skills indicate that usage of figures-of-speech affects the score of an essay greatly [7]. It is interesting to note that there are three characteristics in figures-of-speech which can be used to facilitate CAES to better score Chinese essays. First, it is a writing skill beyond the beginners and requires extensive training. Essays containing figures-of-speech usually score better grades. Second, some subcategories of figures-of-speech are only dependent on both the arrangement of part-of-speech in sentences and the number of words. These subcategories are not related to the semantics of sentences and become feasible to extract. Third, most of figures-of-speech appear in intra-sentence or among few sentences. Hence, the features are not rare.

This paper will focus on “pi-yu” (譬喻) and “pai-bi” (排比), two Chinese figures-of-speech. These two subcategories of figures-of-speech often appear in the essays written by well-disciplined students. The purpose of this paper is to develop and design a method for retrieving the two Chinese figures-of-

speech from essays and to predict the scores of essays based on the two skills used in the essays. A scoring model based on ID3 decision tree algorithm will validate and show that the proposed method increases the performance of CAES effectively.

In the rest of the paper, Section 2 reviews some related studies. Section 3 discusses the proposed method in detail. Section 4 introduces ID3 decision tree algorithm briefly. Section 5 shows the experimental results of the method on real data. Section 6 gives a conclusion and discusses future works.

2 Related Works

Such automated essay scoring systems as PEG[1], IEA[2], TCT[3], e-rater[4], and IntelliMetric[5] for English have been proposed and worked well in the past. Recent studies ([6], [8], [9], and [10]) focus on comparison and evaluation of the performances of these systems. These systems, in general, consist of several modules for retrieving various features from essays. For example, e-rater, which has been used for analytical writing assessment (AWA) in Graduate Management Admissions Test (GMAT), consists of three modules: syntactic, discourse and topic analysis. E-rater extracts features and uses a forward-entry stepwise linear regression method for scoring essays. Another famous AES system Intelligent Essay Assessor (IEA) also comprises three modules: content, style and mechanics. The performances of these systems are satisfactory and close to that of manual scoring. Most of these techniques, however, are not adequate for Chinese essay scoring system due to the lack of grammar parsers for Chinese language.

Ko [7] notes that the usage of figures-of-speech in Chinese essays is an important factor in essay scoring. Many studies ([11], [12], and [13]) have proposed various definitions and classifications for figures-of-speech “pi-yu” and “pai-bi” in Chinese articles. Although definitions are varied, the manifestations of the figures-of-speech are similar to each other. These observations indicate that it is feasible to extract figures-of-speech “pi-yu” and “pai-bi” from essays.

Huang [14] notes that ten syntactic rules for Chinese figures-of-speech are all included in the textbooks of elementary schools, but Chen’s experiments [15] show that the literary “pi-yu” is not used as often as the basic “pi-yu” in sixth grade students’ Chinese essays. Some studies indicate that average students still need to practice the usage of figures-of-speech. Furthermore, [16] and [17] state that students’ writing skills can be enhanced when they practice or study the usage of figures-of-speech.

3 Figures-of-speech “Pi-yu” and “Pai-bi”

Our proposed method is based on two assumptions. First, the writers who use figures-of-speech in essays possess better writing skills. Second, a better writer will use an advanced representation of familiar figure-of-speech skills in essays among many alternatives. Subsection 3.1 discusses the advanced representation of figure-of-speech “pi-yu”. Subsections 3.2 and 3.3 present the method for retrieving “pi-yu” and “pai-bi”, respectively.

Since there is no space between Chinese words, words should first be extracted and tagged. Many studies for Chinese word segmentation and tagging have been proposed. Sinica Autotag will be used in the paper for such preprocessing as word segmentation and part-of-speech tagging. In addition, although comma in Chinese functions as both comma and period in English, the issue of ambiguity does not influence the performance of our method. In brief, this paper treats a Chinese character sequence ended with comma, period, interrogation, exclamation and semicolon as a sentence.

3.1 Building Sets of Connectives and Literary Connectives

Figure-of-speech ”pi-yu” makes a comparison between two unlike elements having at least one quality or characteristics in common. There are mainly four subcategories in “pi-yu”: “ming-yu” (明喻), “an-yu” (暗喻), “jie-yu” (借諭) and “lue-yu” (略喻). “Ming-yu” and “an-yu” comprises three elements:

tenor, connective and vehicle. For example, in sentence “the campus is similar to a market on recess” (下課時校園就像菜市場), words “campus”, “similar” and “market” stand for respectively the tenor, connective and vehicle. “Ming-yu” and “an-yu” are both similar to simile in English, but “ming-yu” differs from “an-yu” in the degree of relationship between tenor and vehicle using different connectives. Because “ming-yu” and “an-yu” occur in essays with specific patterns, this paper only discusses the two subcategories of “pi-yu”.

Connectives are significant identifiers for retrieving the pattern of “pi-yu”. Based on our observations, the parts-of-speech of connectives could be classified into classificatory verbs and conjunctives, respectively denoted as VG and Caa in [18]. For example, words “變成” and “好像”, which are respectively synonymous to word “become” and “like”, are classificatory verbs. Words “跟” (as) and “和” (as) are conjunctives. Since the classificatory verbs and conjunctives contain very few words in Sinica CKIP lexicon, experts can manually select qualified connectives.

Some of the connectives, e.g. word “如” (similar), almost do not appear in low-score essays, but occur in high-score essays frequently. These connectives, denoted as literary connectives, are found to be seldom used in colloquialism. Based on our observations, literary connectives should be useful for essay scoring.

Formula (1) is used to retrieve literary connectives from training data. First essays in training data are divided into a subset of high-score essays and a subset of low-score essays. A literary connective w is defined to satisfy the following condition:

$$\frac{Hf(w)}{Hf(w)+Lf(w)} \geq \beta \tag{1}$$

where $Hf(w)$ represents the numbers of the occurrence of w in the high-score subset, $Lf(w)$ represents that in the low-score subset, β represents a threshold ranged from 0.5 to 1. The higher β value is used, the more discriminatory power the connective has. However, it will result in a small number of literary connectives. Based on our experience from experiments, the best choice of β is 0.6.

3.2 Retrieving Figure-of-speech “Pi-yu”

The appearance of connectives can identify two patterns of figure-of-speech “pi-yu”. The first pattern comprises “noun+connective+noun” in single sentence. For instance, the sentence below:

這時候 學校 變成了 一個 嘈雜的 菜市場
now Campus become a noisy market
(Campus becomes a noisy market now.)

contains the sequence “campus+become+market” which matches the pattern “noun+connective+noun”. Formula (2) describes the rule for the first pattern in detail:

> (Na | Nb | Nca | Ncb) > Connective > (Na | Nb | Nca | Ncb) > (2)

where symbol “>” represents several words or no word, symbol “|” represents logical operator “OR“. Parts-of-speech Na, Nb, Nca, Ncb represent general noun, proper noun, proper place noun and general place noun, respectively.

The second pattern comprises either “connective+adjective+noun” or “connective+noun+adjective” in a single sentence. In addition, it should satisfy two conditions: (i) there is no noun before the connective, (ii) the preceding sentence ends with comma and contains a noun. For example, considering the two adjacent sentences:

校園 充滿 交談的 聲音，就 如 菜市場 般 熱鬧非凡。
Campus fill conversation voice as market boisterous
(Campus fills with conversation voice, just as a boisterous market.)

in which the preceding sentence end with comma and includes noun ”campus”, and the succeeding sentence includes pattern “connective+noun+adjective” corresponding to the sequence “as+market+boisterous” and there is no noun before the connective. Formula (3) describes the rule of the pattern in detail:

$$> \text{Noun} > \text{，} > \text{Connective} > ((\text{Adjective} > \text{Noun}) \mid (\text{Noun} > \text{Adjective})) > \quad (3)$$

where the definitions of symbols “>” and “|” are the same as that in formula (2), “Noun” represents the component (Na | Nb | Nca | Ncb) in formula (2), “Adjective” represents a word whose part-of-speech is denoted as VH or A in [18].

Both rules for “pi-yu” in our proposed method effectively conform to the theoretical structure consisting of tenor, connective, and vehicle. Rule (2) often appears in English sentences and short Chinese sentences. Rule (3) is a mutation of rule (2) where tenor and vehicle appears on different sentences. This is needed because of elaborated description for the tenor and vehicle.

3.3 Retrieving Figure-of-speech “Pai-bi”

Figure-of-speech “pai-bi” uses two sentences or sets of sentences, of which the syntactic structure is similar to each other, to express two concepts of the same property and domain. For example, both sentences “打球 的 打球、散步 的 散步” (Players are playing, walkers are walking.) describe actions in campus using three words and the same syntactic structure: verb following noun. Our proposed method identifies the two single sentences as using the writing skill “pai-bi”.

The following criterion is used to identify if “pai-bi” appears in the essay. If two sentences appearing in a small segment of content contain the same number of words and the same part-of-speech sequence, then the “pai-bi” is considered to occur. For example, in the four serial sentences “到操場走走，可以看到有人悠閒的慢跑；到合作社看看，可以看到有人瘋狂的搶購。” (Some guys are running leisurely on field; some guys are shopping irrationally on snack bar.), the word segmentation and part-of-speech tagging for the first and third sentence is as follows.

到(P) 操場(Ncb) 走走(VA)
到(P) 合作社(Ncb) 看看(VA)

Both sentences consist of three words and the same parts-of-speech of the words. In particular, a preposition, general place noun and verb are in the sequence. Our method hence identifies the occurrence of “pai-bi” in the four serial sentences.

The above example actually shows the delicate aspect of “pai-bi” where the first and the third constitute a usage while the second and fourth also constitute another usage of “pai-bi”. This is an advanced usage of “pai-bi” and is not considered in this study due to its rare occurrence.

4 Scoring with ID3 Decision Tree Algorithm

ID3 algorithm makes use of a set of training examples, which consist of several attributes and a target attribute, to construct a decision tree. The decision tree is used to predict the value of target attribute for a new unseen example. To construct a decision tree, it first searches for an attribute which can best classify the training examples as the root of the decision tree. Next, it creates several branches according to different values of the attribute associated with the root. Each child of the root now contains a different subset of the training examples. Finally, it iteratively constructs a decision tree for the child node. Together, it forms the decision tree of the ID3 algorithm.

Table 1. Examples for constructing ID3 decision tree

	d1	d2	d3	d4	d5	d6	d7	d8	d9
a1	0	0	1	1	0	0	0	1	1
a2	2	1	2	0	0	0	0	1	1
a3	4	3	1	1	2	3	4	4	0
a4	1	0	1	1	0	0	0	0	1
g	H	H	H	L	L	L	L	L	H

It is evident that decision tree algorithm can be used for the application area of scoring essays system. A set of essays can be regarded as the training examples while the score of the essays as the target attribute, the features of essays as other attributes. Table 1 shows 9 essays, denoted as d1, d2, …, d9 with four attributes, denoted as a1, a2, a3 and a4, and a target attribute g. The target attribute g has two values: one is high score denoted as H and the other one is low score denoted as L. Figure 1 shows the decision tree constructed from the training examples of the nine essays.

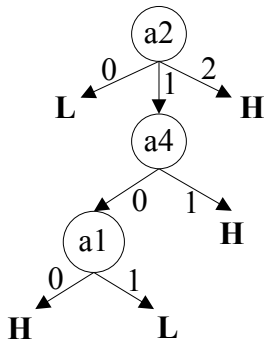

Fig. 1. The decision tree constructed from Table 1

ID3 algorithm uses the entropy measure of a set to select the best attribute for the classification in the algorithm. For details, the readers are referred to [19].

5 Experiments

Experimental corpus consists of 693 essays written by students from eighth grade. The theme of the essays is defined as “Recess at School”. The score of each essay ranges from one point to six point where a higher point represents a higher quality. The score of an essay is obtained by averaging the scores from two or three teachers. The numbers of essays corresponding to different score in the ranges of 1-6 are 40, 166, 234, 177, 70 and 6, respectively. For each experiment, we randomly select three datasets of the size 197, 197 and 299 from the corpus as the training examples, validation examples and testing examples, respectively.

5.1 Relationship Between Score and Figure-of-speech

Table 2 shows how figures-of-speech affect the scores of essays. Row 1 in Table 2 shows the ratios of essays to all of the essays in the corpus under different scores. Row 2 shows the ratios of the essays to all of the essays containing the usage of “pai-bi” under different scores. Row 3 shows the ratios of the essays to all of the essays containing the usage of “pi-yu” under different scores. Row 4 shows the ratios of the essays to all of the essays containing the usage of literary “pi-yu” under different scores. The different distribution or spread of the ratios shows that the usage of figure-of-speech in fact affects the score of the essays.

The total ratios in the expanded column of higher score for row 2, 3 and 4 are 0.57, 0.55 and 0.80 respectively while the total ratios for row 1 are 0.37. It shows that the essays using figure-of-speech increase the odds to obtain higher scores. Further, the data from row 2, 3 and 4 shows that the odds are increased if the advanced skill of figure-of-speech is used. In other words, graders tend to grade essays containing advanced writing skills to higher score against common skills.

Table 2. The distributions of the ratios of essays to all of the essays

	Lower score			Higher score		
	1	2	3	4	5	6
Rates of essays in corpus	0.06	0.24	0.34	0.26	0.10	0.01
Rates of essays containing "pai-bi"	0.02	0.15	0.25	0.42	0.13	0.02
Rates of essays containing "pi-yu"	0.01	0.18	0.26	0.34	0.19	0.02
Rates of essays containing literary "pi-yu"	0.00	0.08	0.13	0.41	0.37	0.02

5.2 Enhancing CAES by Using Figure-of-speech

Table 3 shows that the performance of CAES can be increased by using figure-of speech proposed in our method. The column of "accurate rate" in Table 3 is defined as the ratio of essays in which the score graded by machine and that of the grader is less than one, to all of the essays in the test data. The column of "exact rate" represents the number of the essays, in which the scores graded by machine matches that by human, to all of essays in testing set. The column of "baseline" represents the performance of the decision tree based only on such surface features as the numbers of words, adjectives and idioms.

Table 3. The accurate and exact rates of CAES using different features

Features in essays	Accurate rate		Exact rate	
	baseline	containing feature	baseline	containing feature
"Pai-bi"	0.83	0.83	0.30	0.50
"Pi-yu	0.75	0.83	0.28	0.40
Literary "pi-yu"	0.76	0.85	0.29	0.45
All	0.78	0.91	0.29	0.48

If "pi-yu" is used as a feature of ID3 decision tree, the accurate and exact rates of scoring essays increases 8% and 12%. If literary "pi-yu" is used as a feature of ID3 decision tree, the accurate and exact rates of scoring essays increases 9% and 14%. If literary "pai-bi" is used as a feature of ID3 decision tree, the accurate rate of scoring essays remains steady but the exact rate increases 20%. It shows that any solo usage of figures-of-speech can increase the performance of CAES. Furthermore, if all of three features are used, the accurate and exact rates of scoring essays increase 13% and 19%. The experimental results show that the performance of our proposed method is very close to that of manual grading.

6 Discussions and Future Works

This paper proposed using figures-of-speech "pi-yu" and "pai-bi" as the features of ID3 for scoring essays. It also presents a method for retrieving figures-of-speech "pi-yu" and "pai-bi" from essays. Two advantages are observed in the proposed method. First, the experimental results clearly show that the proposed method increases the accuracy of CAES. Second, the representations of figure-of-speech skills are direct features in essays which human raters also use. Since the direct features represent writer's ability, the scoring model using these features can effectively avoid the bias introduced in surface

features. Moreover, the evaluation of using the direct features in essays can be used as feedback and suggestion for improving writing skills.

Since the patterns identified by the proposed method are not exclusive, further works should be explored in the future. First of all, other patterns of "pi-yu" and "pai-bi" which machine can extract should be examined from the context of linguistics. Moreover, the representation of some figures-of-speech refers to the semantics and concept of words and sentences. It is necessary to develop methods for retrieving semantic patterns of figures-of-speech. Finally, the psychological progress of writers using figures-of-speech should also be analyzed. Such study will not only increase the performance of essay scoring, but also will be useful for developing the e-learning of writing skills.

Acknowledgments. This work was supported in part by the Center for Research on Educational Evaluation and Development, National Taiwan Normal University, Taiwan, under Grant 95E0012-A-05.

References

1. Page, E.B.: Computer Grading of Student Prose Using Modern Concepts and Software. J. of Experimental Education. 67 (1994) 127-142
2. Landauer, T.K., Laham, D., Foltz, P.W.: The Intelligent Essay Assessor. IEEE Intelligent System, 15 (2000) 27-31
3. Larkey, L.S., Croft, W.B.: A Text Categorization Approach to Automated Essay Grading. Automated Essay Scoring: A Cross-Disciplinary Perspective. Lawrence Erlbaum Associates Inc. Mahwah New Jersey (2003) 55-69
4. Burstein, J., Kukich, K., Wolff, S., Lut, C., Chodorow, M., Braden-Harder, L., Dee Harri, M.: Automated Scoring Using A Hybrid Feature Identification Technique. Proc. of the 36th Annual Meeting of the Association of Computational Linguistics. Montreal, Canada. (1998) 206-210
5. Elliot, S.M.: IntelliMetric: From Here to Validity. Proc. of the Annual Meeting of the American Educational Research Association (2001) Seattle, WA, USA
6. Hearst, M.: The Debate on Automated Essay Grading. IEEE Intelligent Systems, 15:5 (2003) 22-37
7. Ko, H.W.: The Evaluation Criteria of Expository and Narrative Writings. J. of Chinese Language Teaching. 1:2 (2004) 15-32
8. Shermis, M.D., Burstein, J.C. (Eds): Automated Essay Scoring: A Cross-Disciplinary Perspective. Lawrence Erlbaum Associates Inc. Mahwah New Jersey (2003)
9. Valenti, S., Neri, F., Cucchiarelli, A.: An overview of current research on automated essay grading. J. of Information Technology Education. 2 (2003) 319-330
10. DIKLI, S.: Automated Essay Scoring. Turkish Online J. of Distance Education 7:1 (2006) 49-62
11. Huang, L.C.: Figure-of-speech "Pai-bi" (I). Chinese Language Monthly. 491 (1998) 19-24
12. Tsai, T.Y.: Rhetoric Theory and Writing Teaching. Newsletter for Teaching the Humanities and Social Sciences 9:3 (1998) 52-62
13. Tsai, M.F.: Figures-of-speech "pi-yu", "bi-ni" and "zhuan-hua", World of Chinese Language and Literature. 16:9 (2001) 81-87
14. Huang, T.T.: Ten Figures-of-Speech Instruction Using Syntax. Elementary Education Century. 215 (2005) 79-88
15. Chen, H.J.: Analysis of Students Use Figure-of-speech "Pi-yu. Newsletter for Languages and Literature Education. 15 (1997) 48-55
16. Chen, C.C.: A Study on Procedural Knowledge Learning for Mandarin Rhetoric Teaching. Master Thesis, Graduate Institute of Taiwan Languages and Language Education, National Hsinchu University of Education, Hsinchu, Taiwan. (2004)
17. Hsu, H.J., Wang, C.C.: The Effect of Reading-and-Writing Rhetoric Instructive Method on the Ability of Elementary Students' Writing Rhetoric Manner. Proc. of Conf. on Children's Languages and Literature Education. Taitung, Taiwan. (1999) 341-362
18. CKIP: Chinese Part-of-speech Analysis. Technical Report 93-05, Academia Sinica, Taipei (1993)
19. Quinlan, J.R.: Induction of Decision Tree, Machine Learning. 1:1 (1986) 81-106