Using LSA and PLSA for text quality analysis

The Case of Autoscoring Essays

Ke Xiaohua
Cisco School of Informatics,
Centre of Translation Studies
Guangdong University of Foreign Studies
Guangzhou, China
carrieke@gdufs.edu.cn

Luo Haijiao*
Cisco School of Informatics
Guangdong University of Foreign Studies
Guangzhou, China
Corresponding author: luohaijiao@gdufs.edu.cn

Abstract— **This paper evaluates the variances of Latent Semantic Analysis (LSA) and Probability Latent Semantic Analysis (PLSA) using EM method on judging text qualities as automated essay (AES) scoring tools. A correlation research design was used to examine the correlation between LSA performance and PLSA performance. Results from the data analyses showed that there was a significant correlation between LSA performance and PLSA performance. Implications of our research for AES reveal that both LSA and PLSA have limited capability at this point and more reliable measures for automated essay analyzing and scoring, like text formats and forms still need to be a part of the text quality analysis.**

Keywords—LSA; PLSA; EM method; AES

I. INTRODUCTION

Latent Semantic Analysis is a statistical latent class model (or aspect model) that has shown satisfy results in scoring content features, such as essays and short-answer responses and related tasks in the last 20 years [1, 2]. New methods such as Probability Latent Semantic Analysis uses probabilistic methods and algebra to search latent space in the corpus is further applied in document clustering. [3] Some of the research found that PLSA was very robust against over fitting for the two tasks, namely LSA and PLSA, considered. On the one hand, aspect models have previously been reported to be very sensitive to over fitting when applied to information retrieval tasks [4] even with small numbers of classes. Therefore, the studies usually proposed using tempered EM in order to avoid overtraining. On the other hand, studies using aspect models for language modeling and parsing report these models to be relatively robust against overtraining when the number of classes is much smaller than the size of the vocabulary [5].

Further insights can be obtained by comparing variance of using LSA and PLSA, especially the performance of autoscoring human essays when adopting different weights calculation methods, so as to offer a promising and challenging research arena for the forthcoming years.

The purpose of this study was to explore and evaluates the scoring performance of LSA and PLSA. Our experiment plotting the number of EM steps vs. different threshold value assigned via applying EM method in PLSA.

II. BUILDING AES MODELS

A. Building LSA model

LSA builds a co-occurrence matrix [A] of words and perform SVD calculation as following

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix} = T \times S \times D'$$

$$= \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1l} \\ T_{21} & T_{22} & \cdots & T_{2l} \\ \cdots & \cdots & \cdots & \cdots \\ T_{m1} & T_{m2} & \cdots & T_{ml} \end{bmatrix} \times \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \sqrt{\lambda_l} \end{bmatrix}$$

$$\times \begin{bmatrix} D_{11} & D_{21} & \cdots & D_{n1} \\ D_{21} & D_{22} & \cdots & D_{n2} \\ \cdots & \cdots & \cdots & \cdots \\ D_{1l} & D_{2l} & \cdots & D_{nl} \end{bmatrix} \tag{1}$$

Matrix A represents the word usage in texts, where each column in [*A*] representing one essay sample and each row representing a unique word-type that appeared in at least two samples. A_{ij} represents the frequency with which the i^{th} word-type appeared in the j^{th} essay sample. To build the LSA scoring model, 3 weights methods: TF*iDF, Entropy, and Cosine can be used before SVD calculation.

B. Building PLSA model

PLSA is a generative model which aims to find a latent topic $Z = \{z_1, \cdots, z_k\}$ from a vocabulary $W = \{W_1, \cdots, W_m\}$ given a set of documents $D = \{D_1, \cdots, D_n\}$ [4]. Z_k indicates a probability relationship between the essay-samples, LSA, and the word-types. PLSA transforms each cell value in $[A^{'}]$ by using formula (2):

$$p(w_j|d_i) = \sum_{k=1}^{k} p(w_j|z_k)\,p(z_k|d_i) \qquad (2)$$

where the probabilities are estimated using EM algorithm that was given in [4], and *k* is defined as hidden-factor and needed an integer value, and $p(w_j|z_k)$ representing the distribution probability of each word-type over all contexts, which is shown in formula (3) and (4):

$$P(z_k|d_i,w_j) = \frac{P(w_j|z_k)\,P(z_k|d_i)}{\sum_{l=1}^{k} P(w_j|z_l)\,P(z_l|d_i)} \qquad (3)$$

$$P(w_j|z_k) = \frac{\sum_{i=1}^{n} a(d_i,w_j)\,P(z_k|d_i,w_j)}{\sum_{j=1}^{m}\sum_{i=1}^{n} a(d_i,w_j)\,P(z_k|d_i,w_j)} \qquad (4)$$

where l is a predefined threshold (in this paper, we assign an integer 10 to l), it decreases with an iterative computation was running, until this formula became convergence. Then, an optimized value of PLSA could be obtained.

III. Experiment

We will focus on the variation of k considering the scoring performance in this paper. We collected 450 essays responding to one same prompt in Chinese language. They were written by native speakers in China. A knowledge base was used to preprocess all texts into word-types. Then, the weights methods were calculated and the LSA or PLSA models were built up for essay scoring.

In order to analyze the scoring performance of the LSA and PLSA models, we designed 5 experiments to build scoring models from LSA and PLSA combining different parameters, like variances of K, l, and a for EM formula. They are listed in Table 1.

TABLE I. Experiments for LSA & PLSA

Experiments	Methods		
	Table column subhead	*K*	*l / alpha*
1	LSA(using TF*iDF, Entropy, Cos)	0-400	/

2	PLSA(using EM)	0-400	l=20,
			a=0.5

3	PLSA(using EM)	0-400	l=40,
			a=0.5

4	PLSA(using EM)	0-400	l=60,
			a=0.5

5	PLSA(using EM)	0-400	rand

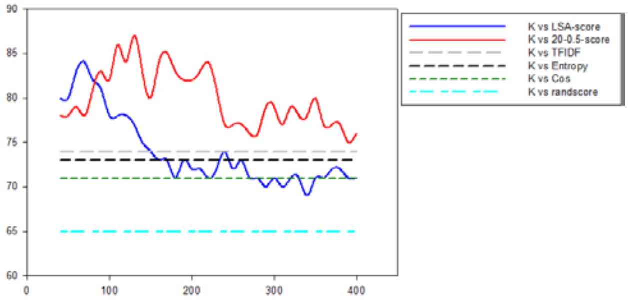
In the standard procedure for LSA, we start by TF*iDF, Entropy, and Cosine methods as experiment 1. PLSA methods with EM formula were applied in experiment 2 through 4. The last experiment used random parameters for EM formula. For saving the calculation time consuming, we adopted a pilot study with smaller data samples (only 50 texts) to provide and brief result for LSA vs PLSA.

IV. Results and Discussion

To investigate the scoring performance of LSA and PLSA models, exact agreement and adjacent agreement between human-machine scores are calculated and compared. Scores are designated as exact if the LSA/PLSA model and human rater agree with each other. Scores are designated as adjacent if LSA/PLSA model has exact + adjacent agreement with the human score.

A. Overall scoring performaces for LSA vs PLSA

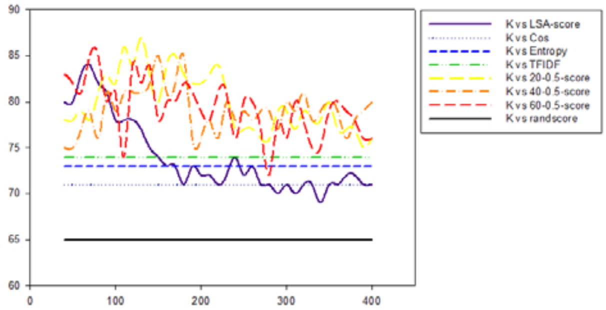
We adopted experiment 1 and experiment 2 to find out the overall scoring performance for LSA vs PLSA with 50 texts. Results are shown in Fig.1.



In Fig.1, red line presents the scoring performance of PLSA with the variance of K and the blue line presents the scoring performance of LSA with the variance of K. These two trend lines showed that, on one hand, the adjacent agreements of LSA and PLSA, with three weight methods, are sensitive with increasing number of K. However, they only increased in a short period, roughly between k value of 50~120. On the other hand, the red line, namely PLSA owns a majority better performance than the blue line, which presents LSA. All these proved that it makes sense to go on with further experiments to find out to what extent could PLSA provide better performance than LSA? This is the main purpose in latter experiments.

B. Comparing LSA and PLSA on the variance of k

We put all the texts in following research. Firstly, we scored texts with LSA method. The human-machine scoring agreement, regarded as the scoring performance is shown in Fig.2 with a purple line.



Then, we applied PLSA with different parameters, like variances of K, l, and a for EM formula to build scoring model and score texts. The parameter K is assigned integers like 20, 40, 60, and rand separately. Results can be found on Fig.2 with colorful dotted lines in yellow, orange, red, and black.

It’s obviously that even for different parameters of PLSA, the overall scoring performance of PLSA is higher than LSA. Most of the trend lines were tend to decrease with the increasing number of K. It appears that the trend lines of LSA and PLSA models allow one to easily capture the peaks of performance from the interval of K. While the good performance of LSA and PLSA are similar, roughly between 0. 83~0.75, again, the best performance can be obtained in PLSA method. But, the peaks appear at different interval of K: for experiment 2, experiment 4, and experiment 4, a good performance can be obtained via [80, 180] , but not for experiment 1, which used LSA method and obtained a best performance via [65, 90]. However, in experiment 6, we cannot fulfill autoscoring task and no performance came with it. Also, there is practically difference in the performance using different K values, more significant are the results for experiment 4 and experiment 2. From the linguistic point of view, one may infer that using PLSA methods could effectively retrieve semantic features from the data set, positively promoting the machine scoring approach. It should be mentioned that, again, the performance of PLSA is just a little better than LSA.

When comparing the Fig.1 and Fig.2, one can easily find out that K value owned minimal impact on scoring performance. They showed that the scoring performance of LSA and PLSA, with different K values as parameters, are sensitive with increasing number of K, though most of they are tend to decrease with the increasing number of K. It appears that the trend lines of LSA and PLSA models allow one to easily capture the peaks of performance from the interval of k. While the good performance of LSA and PLSA are similar, roughly between 0.70~0.85, again, the best performance can be obtained in experiment 2 which used PLSA method.

V. Conclusion

We have applied the LSA and PLSA models to calculate the scores of essays which were pre-scored by human raters written by university students in Chinese language. A correlation research was designed to investigate the scoring performances of LSA/PLSA models considering 3 parameters, i.e., K=20, 40, and 60, in the 6 experiments. From the agreements of the machine scores and human scores, we found that both LSA and PLSA models could be applied to essays scoring systems supplementing human judgment.

The results here have important implications for computer-based educational measurements as well. While more and more automated essay scoring systems may be established, their utility is limited at providing scores, and well-documented assessment strategies like writing portfolios, writing conferences, and the process writing approach should still be included in computer-assisted teaching and learning programs. Besides, more and more reliable measures for assessment, like text formats and forms still need to be a part of the text quality analysis in our future research.

Acknowledgment

The authors would like to express their sincere acknowledgments to the University innovation projects of Guangdong province (2014) as research topics and sponsored this study. The authors also thank for the Guangdong Province Natural Science Fund Project 2015. This manuscript is supported by Innovative School Project in Higher Education of Guangdong, China.

References

[1] S. Dumais, Using semantic analysis to improve access to textual information. Machine Studies, vol.17, pp.87-107, 1982.

[2] Y.Chen, 2012. A topic detection method based on Semantic Dependency Distance and PLSA. The Proceedings of Computer Supported Cooperative Work in Design (CSCWD). 5 2012, pp. 703-708.

[3] X. Ke,Y. Zeng, Q. Ma, L. Zhu. Complex dynamics of text analysis, Physica A: Statistical Mechanics and its Applications (2014),VOL. 415C. pp. 307-314,

[4] T.Hofmann, 1999. Probabilistic Latent Semantic Indexi ng. Proc. SIGIR. 1999.

[5] H.Thomas, 2001. Unsuperised Learning by Probabilistic Latent Semantic Analysis. Machine Learning. 2 2001, pp. 177-196.

[6] X. Ke, Q. Ma. Study on an Impersonal Evaluation System for English-Chinese Translation Based on Semantic Understanding. Perspectives: Studies on Translotology. Vols.22:2(2014.Mar.22), pp.242-254 Taylor & Francis.

[7] J. Burstein, and M. Chodorow, 1999. Automated essay scoring for nonnative English speakers. Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing. 1999.

[8] S. Gui, 2003. The theory of Latent Semantic Analysis and its application. Liguistics and Applied Linguistics. 1 2003, pp. 76-85.

[9] M.D. Shermis, and J. Burstein, 2013. Handbook of Automated Essay Evaluation:Current Applications and New Directions. New York : Routledge, 2013.