Factors Influencing Effectiveness in Automated Essay Scoring with LSA

Fridolin Wild, Christina Stahl, Gerald Stermsek, Yoseba Penya, Gustaf Neumann
Department of Information Systems and New Media,
Vienna University of Economics and Business Administration (WU Wien),
Augasse 2-6, A-1090 Vienna, Austria
{firstname.lastname}@wu-wien.ac.at

Abstract. This paper addresses the ongoing discussion on influencing factors of automatic essay scoring with latent semantic analysis (LSA). Throughout this paper, we contribute to this discussion by presenting evidence for the effects of the parameters text pre-processing, weighting, singular value dimensionality and type of similarity measure on the scoring results. We benchmark this effectiveness by comparing the machine assigned with human assigned scores in a real world case. The paper shows, that each of the identified factors significantly influences the quality of automated essay scoring, but the factors are not to be independent of each other.

Introduction

Computer assisted assessment in education has a long tradition. While early experiments on grading free text responses – run during the Project Essay Grade (PEG) by Page [1] already in 1966 – were carried out on punched cards and had predominantly been syntactical in nature, research today focuses on emulating a human-semantic understanding, backed up by hitherto unknown computing power. Semantic understanding in this case refers to judging the meaning of written essays against them being an exhaustive and satisfying response to an issued question. However, according to Whittington and Hunt [2], free text assessment is a complex and fundamentally subjective process. Several findings report the correlation in grade attribution between two human assessors to be located around 0.6 (cf. e.g. [3]).

Landauer et al. [4] were able to show that by combining the vector model of a document-term-space with singular value decomposition (SVD, a two mode factor analysis) – a method they named 'latent semantic analysis' (LSA) –, grade ranges similar to those awarded by human graders can be produced. Several stages in this process leading from raw input documents to the machine assigned scores allow for optimisations, many of them in common with other application areas of latent semantic analysis and indexing.

The process of auto-scoring can be divided in six sub-steps: *textbase selection*, *text pre-processing*, *weighting*, *calculation of the SVD* and *correlation measurement*. However, contradicting claims can be found concerning the adjustment of influencing factors of these steps.

Perfetti [5], for example, argues for more reliable results with a larger corpus size (input documents). On the other hand, Landauer et al. [6] were able to successfully apply LSA to a corpus with only nine documents.

Nakov [7] reports the best results with a raw term frequency applied as local weighting scheme, whereas Dumais [8] finds a logarithmized term frequency suiting best.

Dimensionality is seen quite different by authors. Dumais [8] sticks to the magical number of 100, whereas Graesser et al. [9] suggest the use of 100 to 300 dimensions. Nakov [7] detects the number of dimensions to vary from 50 to 1500.

Conclusions on how to calibrate an LSA essay scoring process can hardly be drawn from these statements. Furthermore, these examples indicate that identifying the perfect calibration is complex and tightly coupled to the purpose the application serves.

In this contribution we describe and conduct an experiment on varying influence factors and their optimisation within the application of essay scoring using LSA. As a side effect, we provide a proof-of-concept for a real world case with relatively small text-corpora in German. It is not our goal to automatically calculate grades, but to investigate the parameters driving the automated scoring of free text answers with LSA as a basis for automatic feedback and artificial tutoring.

The rest of the paper is structured as follows. In Section 1 we explicate what we understand by essay scoring and expose the algorithm and the setting with which the latent semantic analysis and the scoring process were performed. The methodology applied in this research is documented in Section 2. In Section 3 we explicate our hypothesis and describe the test design of our experiments. The resulting data is analysed in Section 5. Moreover, we review our hypothesis here. Section 5 gives an outlook on future research.

1. Automated Essay Grading with Latent Semantic Analysis (LSA)

1.1 Essays Grading

In opposite to multiple-choice tests, this paper addresses means of how to conduct (auto-) assessment of free text responses with latent semantic analysis (LSA). Hereby the problem of 'feeding' students the right answers, rather than actually testing their active knowledge (cf. Davies), can be avoided.

In accordance with Stalnaker [10] we define an essay to be "… a test item which requires a response composed by the examinee, usually in the form of one or more sentences, of a nature that no single response or pattern of responses can be listed as correct".

Automatic essay scoring furthermore means, that not only a human person skilled or informed in the subject is capable of performing the marking, but – beyond Stalnaker – that a machine can successfully emulate human scoring. In our view this requires an expert skilled or informed in the subject to invest knowledge, and a knowledge engineer to set up a system that is capable of using this represented knowledge to assess incoming essays in respect to their quality and accuracy.

1.2 Latent Semantic Analysis (LSA)

Derived from latent semantic indexing, LSA is intended to enable the analysis of semantic structure of texts [6]. The basic idea behind LSA is that a collocation of terms of a given document-term-space reflects a higher-order – latent semantic – structure, which is obscured by word usage (e.g. synonyms or ambiguities). By using conceptual indices that are derived statistically via a truncated singular value decomposition, this variability problem can be overcome (cf. [11]).

Figure 1. Singular Value Decomposition (original left, truncated right).

A document-term-matrix is constructed from a given text base of n documents containing m terms (see M in figure 1). This matrix M of the size m×n is then resolved by the singular value decomposition into the term vector matrix T (constituting the right singular vectors), the document vector matrix D (constituting the right singular vectors) being both orthonormal and the diagonal matrix S. Multiplying the truncated matrices T_k, S_k and D_k results in a new matrix M_k which is the least-squares best fit approximation of M with k singular values.

In case column 1 of M_k constitutes the document vector of a perfect essay (a 'golden' essay) and column 6 of M_k were an essay to be graded, the grade could be calculated as the correlation between these two vectors.

To keep the essays that are to be tested from influencing the factor distribution, they can be folded-in afterwards by calculating their pseudo document vector m_i in M_k as follows in equation 1 and 2 (cf. [11]):

$$d_i = v^T T_k S_k^{-1} \quad (1)$$

$$m_i = T_k S_k d_i^T \quad (2)$$

The document vector v^T in equation 1 is identical to an additional column of M with the term frequencies of the essay to be folded-in. T_k and S_k are the truncated matrices from an SVD applied on a given text collection used to construct the latent semantic space. For essay scoring the text collection usually consists of text book sections, model answers, glossary entries, generic texts, and the like (cf. [6], [9]).

2. Methodology

Automatic essay scoring with a software programme allows us to alter the settings of the suspected influencing factors, we adopted for an experimental approach. An experiment tries to explore the cause-and-effect relationship where causes can be manipulated to produce different kinds of effects [12].

For our experiments we used a real world text corpus of students' answers to a marketing exam question. The corpus consisted of 43 files, which were pre-graded by a human assessor with points from 0 to 5. We assumed that every point is of the same value. So, when summed up, the resulting scale of the total scores is equidistant in its value representation.

The essays were in average 56.4 words long. From the text corpus three 'golden essays' were taken to compute the correlation for the remaining essays. To build the SVD we used a marketing glossary consisting of 302 files, each file containing one glossary entry. The glossary is part of the course material offered via our e-learning application learn@WU of our university. The average length of the glossary entries was 56.1 words.

This enabled us to compare machine assigned scores (our dependent variables) to the human assessed scores by measuring their correlation. By changing consecutively and ceteris paribus the influencing factors (our independent variables) we investigated their influence on the score correlation.

3. Hypothesis and Test Design

We conducted several tests addressing four aspects that have proven to show great influence on the functionality and effectiveness of LSA: (1) the pre-processing of the input text, (2) the use of weighting-schemes, (3), the choice of dimensionality, and (4) the applied similarity measure (see figure 2).

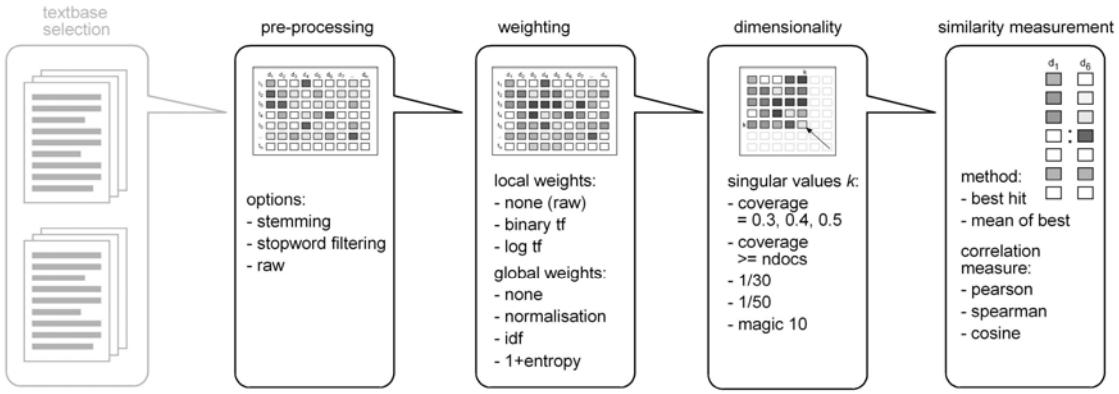


Figure 2. Considered influencing factors.

3.1 Document pre-processing

Document pre-processing is a common procedure in information retrieval and comprises several text operations such as lexical analysis, use of stop-word lists, stemming, selection of index terms, and construction of thesauri [13].

For our purpose we focused on the elimination of stop-words and stemming. We used a stop-word list with 373 German terms. As stemmer we applied Porter's Snowball stemmer. Although research shows that document pre-processing in general yields better results for information retrieval this is not necessarily true for LSA. In their experiments Nakov et al. [14] obtained better overall results when removing stop-words but a significant improvement for stemming was only found in one case. This supports the hypotheses that the removal of stop-words improves LSA, whereas the use of stemming-algorithms does not.

The effects of pre-processing were assessed by testing the corpus with and without stemming as well as with and without stop-word removal. Furthermore, we tested the combination of stemming and stop-word removal. For the succeeding tests, we used the raw matrix as default.

3.2 Weighting-Schemes

Weighting-schemes have shown to have the most extensive influence on the effectiveness of LSA. Several weighting-schemes – both local and global – have been tested for various languages and applications of LSA yielding best results for the logarithm as the local, and the entropy as the global weighting [e.g. 8, 15]. We have every reason to assume that these results will also hold true for the German language and for the automated scoring of essays.

For testing the influence of weighting-schemes on LSA we combined three local (raw term-frequency, logarithm, and binary) and three global (normalization, inverse document-frequency, and entropy) weightings. We also performed tests without any global weighting, leading to 3 x 4 = 12 test runs. As default we used the raw term frequency.

3.3 Choice of Dimensionality

Among other factors the choice of dimensionality has a significant impact on the results of any distance measurement conducted on the vector space of the underlying data. After the application of the singular value decomposition on the original term-document matrix, a reduced matrix it is reconstructed using only the k-largest singular values. This aims at an approximation to the original vector space by a reduction of the dimensionality. This approach captures the most important structure inherent in the original matrix, reducing noise and variability in word usage [11]. In this paper we consider the following possibilities to determine to number of selected factors in the approximated vector space:

- *Percentage of cumulated singular values*: Using a normalized vector of cumulated singular values we can sum up singular values until we reach a specific value. In our paper we suggest to use 50%, 40% and 30% of the cumulative summed up singular values.
- *Absolute value of cumulated singular values equals number of documents*: With this approach the sum of the first k singular values factors equals the number of documents n used in the analysis.
- *Percentage of number of terms*: Alternatively the number of used factors can be determined by a fraction of used terms. Typical fractions are 1/30 or 1/50.
- *Fixed number of factors*: A less sophisticated and inflexible approach is to use a fixed number of factors, e.g. 10 factors. The number has to be determined depending on the underlying text corpus.

With the variation of dimensionality of the reduced vector space we expect to gain information on which approach of choosing the dimensionality fits best our requirements in the realm of automated test assessment. As a standard value we will be using the fixed number of 10 factors, which we expect to work out best.

3.4 Similarity Measures

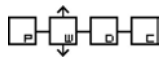
Finally, we tested three similarity measures: the Pearson-correlation, Spearman's r and the cosine. As the cosine is the most commonly used measure for comparing LSA-vectors and usually works best for information retrieval [16] we also expect it to yield the best results.

4. Reporting Results

The results of our test series can be seen in the following tables 1–4. As we expected, stop-word removal proved best (see table 1), but – against our hypothesis – results can be slightly improved further by combining it with stemming. However, stemming alone reduced the performance in scoring correlations with the human assessed scores. With the parameters of the remaining influencing factors weighting, dimensionality and correlation measurement being set to the default values as stated above, the resulting scores of stop-word removal and the combination of stopping and stemming correlate on level below 0.05 significantly with the human scores.

For the next set of parameters in the weighting step we assumed, that as widely cited in the literature, logarithmized term frequencies as local weights multiplied by entropy would produce the best results. Compared to taking raw or binary frequencies as local weight, the logarithm seems to have hardly any influence at all. It produces slightly worse results than using the raw matrix, but minimally outperforms binary frequencies. However, all three local weights alone do not produce a significant correlation with the human scores. Within the

global weighting schemes, the inverse document frequency shows outstanding results. Combined with the raw or logarithmized document frequency, IDF is even on a level below 0.01 significantly correlated with the human scores. Normalisation seems to have nearly no influence and was never correlated significantly. Contrary to our hypothesis, entropy reduced the performance in every case.

		Kendall-Tau-b		Spearman-Rho	
		mean	max	mean	max
raw	*cor*	0.079	0.118	0.094	0.151
	sig	0.51	0.325	0.548	0.333
raw × norm	*cor*	0.162	0.193	0.203	0.242
	sig	0.176	0.108	0.192	0.118
raw × idf	*cor*	.379(**)	.349(**)	.474(**)	.441(**)
	sig	0.002	0.004	0.001	0.003
raw × entropy	*cor*	0.035	0.068	0.036	0.101
	sig	0.772	0.571	0.821	0.52
log	*cor*	0.074	0.099	0.101	0.139
	sig	0.54	0.411	0.521	0.372
log × norm	*cor*	0.162	0.193	0.207	0.244
	sig	0.176	0.108	0.182	0.115
log × idf	*cor*	.304(*)	.357(**)	.392(**)	.453(**)
	sig	0.011	0.003	0.009	0.002
log × entropy	*cor*	0.06	0.085	0.074	0.123
	sig	0.619	0.48	0.636	0.432
bintf	*cor*	0.068	0.135	0.089	0.164
	sig	0.571	0.262	0.569	0.295
bintf × norm	*cor*	0.173	0.207	0.225	0.257
	sig	0.148	0.085	0.148	0.097
bintf × idf	*cor*	.273(*)	.302(*)	.360(*)	.388(*)
	sig	0.023	0.012	0.018	0.01
bintf × entropy	*cor*	0.012	0.065	0.03	0.091
	sig	0.917	0.587	0.851	0.562

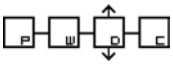
			Kendall-Tau-b		Spearman-Rho	
			mean	max	mean	max
share = 0.5	**81**	*cor*	.359(**)	.329(**)	.436(**)	.408(**)
		sig	0.003	0.006	0.003	0.007
share = 0.4	**56**	*cor*	.365(**)	.332(**)	.448(**)	.419(**)
		sig	0.002	0.006	0.003	0.005
share = 0.3	**35**	*cor*	.334(**)	.293(*)	.407(**)	.363(*)
		sig	0.005	0.015	0.007	0.017
ndocs	**9**	*cor*	0.09	0.11	0.109	0.157
		sig	0.452	0.361	0.487	0.316
1/30	**10**	*cor*	0.079	0.118	0.094	0.151
		sig	0.51	0.325	0.548	0.333
1/50	**6**	*cor*	0.054	0.129	0.053	0.185
		sig	0.652	0.282	0.736	0.235
fixed = 10	**10**	*cor*	0.079	0.118	0.094	0.151
		sig	0.51	0.325	0.548	0.333

Table 3. Correlation between machine assigned and human assigned scores for each dimensionality calculus.

*) Correlation is significant on a level smaller than 0.05

**) Correlation is significant on a level smaller than 0.01

Table 2. Correlation of machine assigned scores with human scores for each weighting scheme.

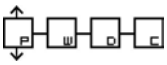
		Kendall-Tau-b		Spearman-Rho	
		mean	max	mean	max
raw	*cor*	0.079	0.118	0.094	0.151
	sig	0.255	0.163	0.274	0.167
stemmed	*cor*	0.035	0.057	0.034	0.083
	sig	0.386	0.318	0.416	0.298
stopped	*cor*	.215(*)	0.152	.282(*)	0.192
	sig	0.036	0.104	0.034	0.109
stemmed & stopped	*cor*	.246(*)	.202(*)	.304(*)	.258(*)
	sig	0.02	0.047	0.024	0.047

Table 1. Correlation of machine and human scores with varying pre-processing methods.

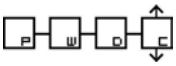
		Kendall-Tau-b		Spearman-Rho	
		mean	max	mean	max
pearson	*cor*	0.079	0.118	0.094	0.151
	sig	0.51	0.325	0.548	0.333
cosine	*cor*	0.079	0.118	0.094	0.15
	sig	0.51	0.325	0.548	0.336
spearman	*cor*	.248(*)	.316(**)	.312(*)	.406(**)
	sig	0.038	0.009	0.042	0.007

Table 4. Correlation for different similarity measures.

*) Correlation is significant on a level smaller than 0.05

**) Correlation is significant on a level smaller than 0.01

Falling short of our expectations, we found evidence in our seven tests about dimensionality, that calculating the number of dimensions as a share of the normalized and cumulated singular values helps to increase the effectiveness significantly on a level below 0.01, whereas the other methods failed to show significant influence: summing up singular values until they

equal the number of documents, taking 1/30 or 1/50 of the number of singular values, and using the fixed number of 10 seem to be fairly bad estimations of a good number of dimensions to be applied. The share of 40% slightly outperforms the others.

Comparing the influence on effective correlation to the human scores of different similarity measures, choosing spearman's rho worked out best. It was the only measure producing a correlation on a level below 0.01 with the human scores.

5. Conclusions and Future Work

In this paper we investigated the influencing factors on effectiveness in automatic essay scoring. Our results give evidence, that for the real world case we tested, the identified parameters drive the correlation of the machine assigned with the human scores. However, several recommendations on the adjustment of these parameters, which we extracted from literature, do not apply in our case. In fact, optimisations seem not to be independent of each other. Furthermore, we suspect that their adjustment strongly relies on the document corpus used as text base and on the essays to be assessed.

Nevertheless, significant correlations between machine and human scores could be discovered, which ensures, that the applied LSA method can be exploited to automatically create valuable feedback on learning success and knowledge acquisition.

Based on these first results, we intend to test the dependency of the parameter settings on each other for all possible combinations. Additionally, the stability of the results within the same discipline and in different contexts needs to be further examined. Furthermore, we intend to investigate scoring of essays not against best-practice texts, but against single aspects, as this would allow us to generate a more detailed feedback on the content of essays.

References

[1] Page, E. (1966). The imminence of grading essays by computer. Phi Delta Kappan, 47, 238-243.

[2] Whittington, D., Hunt, H. (1999): Approaches to the computerized assessment of free text responses, Proceedings of the 3rd CAA Conference, Loughborough.

[3] Landauer, T., Psotka, J. (2000): Simulating Text Understanding for Educational Applications with Latent Semantic Analysis: Introduction to LSA, In: Interactive Learning Environments, vol. 8, no. 2, pp. 73-86(14)

[4] Landauer, T., Foltz, P., Laham, D. (1998): Introduction to Latent Semantic Analysis, In: Discourse Processes, 25, pp. 259-284

[5] Perfetti, C. (1998): The Limits of Co-Occurrence, In: Discourse Processes, 25 (2&3), pp. 363-377

[6] Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R. (1990): Indexing by Latent Semantic Analysis, In: Journal of the American Society for Information Science, 41(6), pp. 391-407

[7] Nakov P. (2000): Getting Better Results with Latent Semantic Indexing, In: Proceedings of the Students Presentations at the European Summer School in Logic Language and Information (ESSLLI'00), pp. 156-166

[8] Dumais, S. (1990): Enhancing Performance in Latent Semantic Indexing (LSI) Retrieval. Technical Report, Bellcore

[9] Graeser, A., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R. (1999): AutoTutor: A simulation of a human tutor, In: Journal of Cognitive Systems Research, 1, pp. 35-51

[10] Stalnaker, J. M. (1951). The Essay Type of Examination. In E. F. Lindquist (Ed.), Educational Measurement (pp. 495-530). Menasha, Wisconsin: George Banta.

[11] Berry, M., Dumais, S., O'Brien, G. (1995): Using Linear Algebra for Intelligent Information Retrieval, In: SIAM Review, Vol. 37(4), pp. 573-595

[12] Picciano, A. (2004): Educational Research Primer. Continuum, London.

[13] Baeza-Yates, R., Ribeiro-Neto, B. (1999): Modern Information Retrieval. ACM Press, New York.

[14] Nakov, P., Popova, A., Mateev, P. (2001): Weight functions impact on LSA performance. In: Recent Advances in Natural language processing – RANLP'2001. Tzigov Chark, Bulgaria, pp. 187-193.

[15] Nakov, P., Valchanova, E., Angelova, G. (2003): Towards Deeper Understanding of the LSA Performance. In: Recent Advances in Natural language processing – RANLP'2003, pp. 311-318.

[16] Landauer, T., Dumais, S. (1997): A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. In: Psychological Review 104 (1997) 2, pp. 211-240.