Aplicação de um método LSA na avaliação automática de respostas discursivas

João Carlos Alves dos Santos[1], Tácio Ribeiro[2], Eloi Favero[2], Joaquim Queiroz[3]

[1]Faculdade de Matemática – Universidade Federal do Pará(UFPA)
– Belém – PA – Brazil

[2]Programa de Pós-Graduação em Ciência da Computação – Universidade Federal do Pará
– Belém – PA – Brazil

[3]Faculdade de Estatística
Universidade Fedral do Pará (UFPA) – Belém, PA – Brazil

{jcas,tacio,favero,joaquim}@ufpa.br

Abstract. *In order to attend the virtual learning environment needs, this paper presents the LSA (Latent Semantic Analysis) application to estimate scores automatically in open ended questions, because still there is not a method with a acceptable accuracy for practical uses. It was developed a model based in LSA composed of six steps: pre-processing, normalization, weighting, singular value decomposition, classification and model fitting. As result obtained a study about the LSA parameters adjustment for practical uses in automatic assessment of discursive responses, and as best result was obtained a accuracy of 87,35%.*

Resumo. *Para atender necessidades dos ambientes virtuais de ensino, este trabalho apresenta a aplicação da Análise Semântica Latente (LSA) para estimar as pontuações obtidas em respostas a questões discursivas, pois ainda não se tem um método com acurácia aceitável para uso prático. Foi desenvolvido um modelo baseado em LSA formado por seis etapas: pré-processamento, normalização, ponderação, decomposição a valores singulares, classificação e ajustes do modelo. Como resultado, obteve-se um estudo sobre o ajuste de parâmetros do método (LSA) para uso prático em avaliações automáticas de respostas discursivas e o melhor resultado obtido foi uma acurácia de 87,35%.*

1. Introdução

Durante sua escolarização, o aluno passa naturalmente por um processo de avaliação de ensino aprendizagem contínuo, cumulativo e sistemático. Mesmo diante das concepções pedagógicas mais modernas, a aplicação de avaliações com questões do tipo discursivas tem grande relevância, pois elas avaliam a capacidade de leitura, interpretação e construção do texto do aluno. No entanto, a tarefa de correção manual de respostas discursivas ou ensaios demanda muito tempo do professor.

O desenvolvimento de ferramentas que automatizem a correção de respostas discursivas, em ambientes virtuais de ensino, traz vantagens para professores e alunos: 1) feedback imediato para os alunos guiando, reorientando e estimulando seus próximos passos; 2) permite o um ranking continuo da turma, tornando o aluno ciente da sua avaliação em

relação a seus colegas; 3) libera a carga de trabalho do professor com a correção manual dessas respostas; 4) permite ao professor um continuo acompanhamento da performance da turma, com a identificação de situações extremas, onde ele deverá focar seus esforços ou redirecionar o programa. Apesar da relevância deste [Valenti et al. 2003] ainda não existem trabalhos com acurácia aceitável para aplicação prática nos atuais sistemas virtuais de ensino.

No sentido de contribuição, esta pesquisa foca no desenvolvimento de uma ferramenta de avaliação automática com a abordagem LSA. Esta abordagem é promissora na avaliação automática de respostas discursivas, entretanto sua eficácia depende fortemente do ajuste de seus parâmetros e do domínio de aplicação [Lifchitz et al. 2009]. A idéia é que o método LSA faça uma estimativa das pontuações obtidas por alunos em respostas discursivas de tal modo que a acurácia desta estimativa seja minimamente aceitável para uso prático do método. O corpus[1] da pesquisa foi constituído por respostas a duas questões discursivas, uma de natureza discursivo-conceitual e outra discurvivo-argumentativa, que foram avaliadas previamente por dois avaliadores humanos. Representamos o corpus de resposta por uma matriz onde o número de linhas representa o vocabulário das palavras e cada coluna é a representação vetorial de cada resposta. Propomos um método LSA com uma arquitetura em seis etapas [Wild et al. 2005]: 1) pré-processamento, 2) normalização 3) ponderação, 4) decomposição a valores singulares, 5) medidas de similaridade e 6) ajustes do modelo. Considerou-se o ajuste de parâmetros dentro de cada uma destas etapas.

Na etapa de pré-processamento que representa o corpus com a variação dos métodos de unigramas e bigramas de palavras em conjunto com as técnicas de remoção de stop words[2] e de stemming[3] . Na etapa de normalização substituímos cada vetor por ele mesmo dividido pela sua própria norma[4]. Na etapa de ponderação aplicamos uma função peso que estima a importância de cada palavra no corpus. Em seguida obtemos a decomposição a valores singulares da matriz bem como escolhemos a dimensão do espaço semântico. Na etapa de classificação, verificamos a similaridade entre as respostas usando como medidas de similaridade: cosseno, correlação de Pearson, correlação de Spearman, e distância de Minkowski. Na etapa de ajustes do modelo, com o intuito de corrigir distorções, aplicamos um fator de penalização à nota atribuída a uma resposta que contém um número de palavras abaixo da média menos o desvio padrão.

Executamos inúmeras iterações do processo todo, guardando as melhores configurações de cada experimento realizado. Como resultado dos procedimentos, obteve-se 87,35% como melhor índice de acurácia. Além desta introdução, têm-se mais quatro seções, onde a seção 2 faz uma breve descrição sobre LSA; a seção 3 descreve o corpus usado nessa pesquisa; a seção 4 descreve o método proposto; a seção 5 apresenta a metodologia da

[1]Um conjunto de textos escritos ou falados numa língua, disponível para análise

[2]São palavras que podem ser consideradas irrelevantes para o conjunto de resultados a ser exibido em uma busca realizada por um sistema de busca

[3]Radicalização ou determinação do radical visa reduzir as variações de uma mesma raiz vocabular

[4]Consideramos a norma euclidiana $\| \cdot \|_2$

pesquisa ; a seção 6 apresenta os resultados obtidos e por fim, a seção 7 apresenta as conclusões e trabalhos futuros.

2. Breve descrição sobre LSA

Latent Semantic Analysis (LSA) é uma técnica estatístico-matemática de extração e inferência de relações do uso contextual de palavras em passagens de um texto em um corpus. O primeiro passo é representar o corpus por uma matriz termo-documento, que denotaremos por A, do tipo $m \times n$, onde m representa o número de palavras distintas e n o número de textos que compõe o corpus. Cada entrada da matriz A é ponderada por uma função peso que estima a importância de cada palavra no texto em que ela está contida e seu grau de influência como um todo. Em seguida, calcula-se a decomposição a valores singulares (SVD) da matriz termo-documento, o que revela a arquitetura das correlações entre as palavras nos textos. Esta decomposição produz uma fatoração de A como sendo um produto de três matrizes da seguinte maneira: $A = USV^t$, onde U e V são matrizes ortogonais quadradas de ordens m e n, respectivamente, e $S = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix}$, sendo D uma matriz diagonal cujos elementos diagonais são os r valores singulares de A, onde r é o posto[5] da matriz A. Após a decomposição aproximamos a matriz A por outra matriz. Esta aproximação é feita da seguinte maneira: escolhemos as k primeiras linhas e colunas das matrizes U, S e V, e construímos a matriz $A_k = U_k S_k V_k^t$. A razão desta escolha é o fato de que nas primeiras colunas dessas matrizes estão os autovetores associados aos valores singulares de maior módulo. A matriz aproximação A_k, tem as mesmas dimensões da matriz A e tem posto k, bem menor do que r. A matriz A_k é a melhor aproximação por mínimos quadrados da matriz A [Deerwester et al. 1990]. Assim, LSA transfere a discussão do espaço inicial para um espaço de dimensão bem menor que é o espaço gerado pelas colunas da matriz A_k, chamado espaço semântico, sendo que neste espaço se realiza a etapa de classificação entre os textos. As coleções de textos usualmente são seções de livros, respostas modelos, entrada de glossários, textos genéricos e similares.

3. O corpus da Pesquisa

O corpus da pesquisa foi constituído por respostas a duas questões discursivas: uma conceitual e outra discursiva. Estas duas questões fazem parte do boletim de questões da terceira fase do Processo Seletivo Seriado 2008 da Universidade Federal do Pará. O boletim era composto por questões discursivas das seguintes disciplinas: Biologia, Física, Geografia, Língua Estrangeira, Língua Portuguesa, Literatura, Matemática e Química. O total de questões foi de 26, sendo duas de Língua Portuguesa e três de cada uma das demais disciplinas, sendo que o aluno deveria responder apenas uma questão por cada disciplina.

Para composição do corpus, de um universo de 1000 folhas de respostas, disponibilizados pelo órgão da UFPa responsável pelo processo seletivo, foram selecionadas aleatoriamente 130 folhas respostas de uma questão de Biologia e 229 folhas respostas de uma questão de Geografia. A questão da área de Biologia propunha a elaboração de três conceitos de uma dada taxonomia da Citologia e a questão da área de Geografia propunha

[5]Posto de uma matriz é a dimensão do espaço das linhas (ou colunas) da própria matriz

a elaboração de uma argumentação em defesa de dado ponto de vista formado a respeito da Geografia Humana e Econômica da Região. Para viabilização dessa pesquisa foi necessário digitar manualmente cada resposta, a fim de que estas pudessem analisadas pelo método proposto. Durante este processo de entrada de dados foram feitas apenas correções de ortografia sem alterar a concordância gramatical do texto original.

4. O Método Proposto

O método LSA proposto pressupõe que o corpus da pesquisa seja representado por uma matriz, a qual é freqüentemente chamada de termo-documento. Para gerar esta matriz codificamos um programa que fez a contagem das palavras individualmente (unigramas) ou a contagem de seqüências de duas palavras (bigramas) das respostas. Codificamos as linhas como sendo a quantidade de ocorrências de unigramas (ou bigramas) e as colunas como as respostas, sendo que a resposta base foi codificada como a primeira coluna. O método estima a pontuação para cada resposta através de uma arquitetura LSA dividida em seis etapas: 1) pré-processamento das respostas, 2) normalização, 3) ponderação, 4) decomposição a valores singulares, 5) Classificação e 6) ajustes do método. Comparamos as pontuações estimadas pelo método com as pontuações por avaliadores humanos calculando a acurácia entre elas. Os procedimentos do método LSA proposto são descritos pelo algoritmo abaixo:

1. Construção e leitura da matriz termo-documento;
2. Ponderação de cada entrada da matriz através de uma transformação peso que expressa a importância das palavras nas respostas;
3. Calcula a SVD da matriz;
4. Reduz para o espaço semântico;
5. Classifica as respostas;
6. Calcula a média e o desvio padrão do número de palavras por resposta;
7. Aplica um fator de penalização;
8. Reclassifica os ensaios;
9. Calcula o erro cometido;
10. Calcula a acurácia;
11. Repete os passos acima alterando os parâmetros e guarda a melhor configuração encontrada.

5. Metodologia

A arquitetura utilizada possibilitou uma abordagem empírica que permitiu encontrar o melhor ajuste de parâmetros que influenciam na eficiência do método. Executou-se repetidamente as etapas da arquitetura do método proposto para todas as combinações possíveis, a fim de se encontrar a melhor configuração para obtenção da melhor acurácia da estimativa contra avaliadores humanos. Em cada etapa ajustou-se sucessivamente fatores que influenciam na eficácia do método.

Na etapa de pré-processamento construímos a matriz termo-documento através das abordagens de unigramas e bigramas das respostas combinando com as técnicas de remoção de stop word e stemmimg, o que possibilitou a construção de seis matrizes termo-documentos para cada um dos conjuntos de respostas. No caso da abordagem

bigramas, para evitar uma grande quantidade de zeros, contamos apenas aqueles bigramas que aparecem em pelo menos duas respostas. Para a questão de Biologia codificamos a resposta base através de um texto, dado por um especialista humano, com todos os conceitos corretos para as escolhas possíveis nas respostas. Na codificação da resposta base para a questão de Geografia utilizamos outra abordagem: concatenamos cinco respostas dos alunos que obtiveram a maior pontuação por avaliadores humanos.

Na etapa de normalização consideramos a possibilidade de considerar as colunas da matriz termo-documento como vetores unitários.

A tabela 1 mostra os esquemas de ponderação utilizados na etapa de ponderação:

Tabela 1. Ponderações locais e globais

Ponderações locais			
	termo frequência	$ft(i,j) = \frac{c(i,j)}{\sum_{k=1}^{I} c(k,j)}$	$c(i,j)$ é o número de vezes que a palavra i ocorre no ensaio j e I é o número total de palavras
	binária	$\begin{cases} 1, & \text{se } ft(i,j) > 0 \\ 0, & \text{se } ft(i,j) = 0 \end{cases}$	
	logaritmo	$\log_2(ft(i,j)+1)$	
	$\frac{norma\ euclidiana}{soma\ componentes}$ 6	$\frac{\|t_j\|_2}{\sum_{k=1}^{I} c(k,j)}$	t_j é o $j-$ ésimo vetor coluna
Ponderações globais			
	entropia	$1 + \frac{\sum_{k=1}^{n^o\ ensaios} p(i,k) \cdot \log_2 p(i,k)}{\log_2 n^o\ ensaios}$	$p(i,k) = \frac{ft(i,k)}{gf(i)}$ é a probabilidade condicional e $gf(i)$ é a frequência global da palavra i
	frequência documento inverso (idf)	$1 + \log_2(\frac{n^o\ ensaios}{df(i)})$	$df(i)$ é o número de ensaios em que a palavra i aparece
	normal	$\frac{1}{\sqrt{\sum_{k=1}^{n^o\ ensaios} (local(i,k))^2}}$	$local(i,k)$ é uma ponderaçao local

Tabela 2. Medidas de Similaridade

medidas de similaridade	
cosseno	$\frac{t_j \cdot t_k}{\|t_j\|_2 \|t_k\|_2}$
correlação de Pearson	$\frac{cov(t_j, t_k)}{\sqrt{var(t_j) \cdot var(t_{k)})}}$
correlação de Spearman	$1 - \frac{6\sum_{i=1}^{I} d_i^2}{I^3 - I}$
distância de Minkowski	$\sqrt[q]{\sum_{i=1}^{I} d_i^q}$

É na etapa de decomposição a valores singulares da matriz termo-documento que escolhemos a dimensão k para o espaço semântico. Esta etapa tem impacto significativo sobre os resultados obtidos pelo método, pois é neste espaço que se classifica as comparações entre a resposta base e as demais respostas [Nakov 2000]. Na literatura existem algumas sugestões para escolha do número k, dentre elas optamos por variar k de 1 até o número total de respostas e escolhendo aquele que forneceu a melhor classificação.

Na etapa de classificação utilizamos cinco medidas de similaridades: cosseno, coeficiente de correlação de Pearson, coeficiente de correlação rho de Spearman e duas distâncias de Minkowski, em particular a distância euclidiana. Na Tabela 2 temos as fórmulas que definem cada uma destas similaridades, onde $t_j = (c_1 j, \cdots, c_I j)$ é a representação vetorial de uma resposta, sendo j o número de respostas e I o número total de palavras.

Durante a etapa de classificação notamos grandes distorções entre as pontuações atribuídas pelo método proposto e por avaliadores humanos para respostas com um número pequeno de palavras. Verificamos as pontuações obtidas pelos alunos em função do número de palavras por respostas. A média de palavras das 130 respostas para a questão de Biologia foi de 27,84, com desvio padrão de 17,11 e a média de palavras das 229 respostas para a questão de Geografia foi de 67,65, com desvio padrão de 30,25. A Figura 1 mostra as pontuações obtidas pelas respostas por avaliadores humanos para as questões de Biologia e Geografia.

Para as respostas de Biologia verifica-se que respostas com poucas palavras obtiveram baixa pontuação, enquanto que respostas com pontuação máxima foram aqueles com número de palavras acima da média menos desvio padrão. Para as respostas de Geografia verifica-se que respostas com um número baixo de palavras obtiveram uma pontuação baixa e respostas com número de palavras acima da média mais o desvio padrão tiveram uma pontuação mais elevada.

Na etapa de ajuste do método implementou-se um fator de penalização para corrigir discrepâncias entre as pontuações humana e automática de respostas com um número de pa-

Figura 1. Pontuações por número de palavras

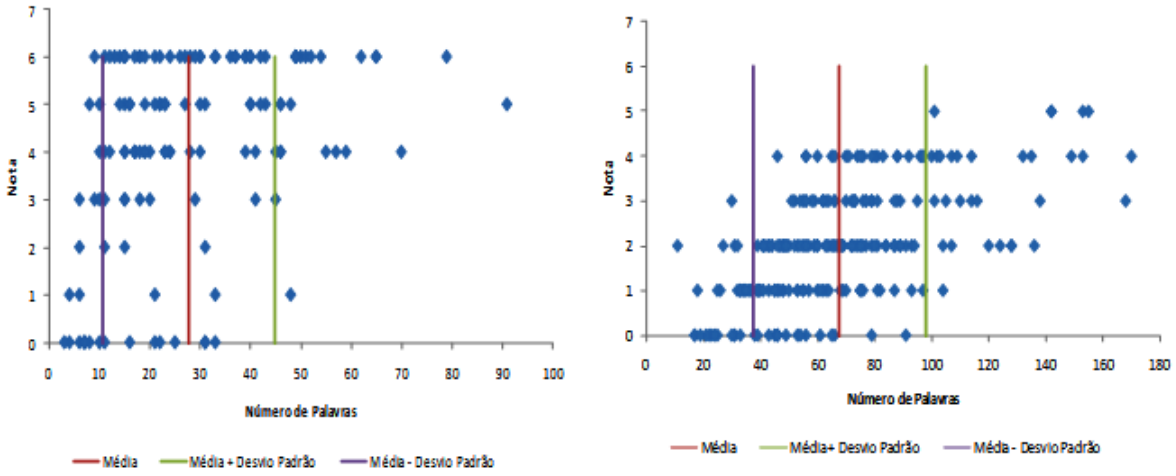


Tabela 3. Melhores parâmetros

	pré-processamento	ponderaçao local	ponderaçao global	dimensão k	similaridade	acurácia
Respostas Biologia	unigrama	termo frequência	IDF	4	pearson	84,24
	bigrama	norma euclidiana/soma componentes	entropia	6	cosseno	84,78
Respostas Geografia	unigrama	Binária	entropia	3	pearson	86,9
	bigrama	norma euclidiana/soma componentes	-	6	cosseno	84,78

lavras abaixo da média menos o desvio padrão. Em seguida reclassificamos as pontuações obtidas pelas respostas, calculamos o erro cometido e a acurácia da estimativa.

6. Resultados obtidos nos experimentos

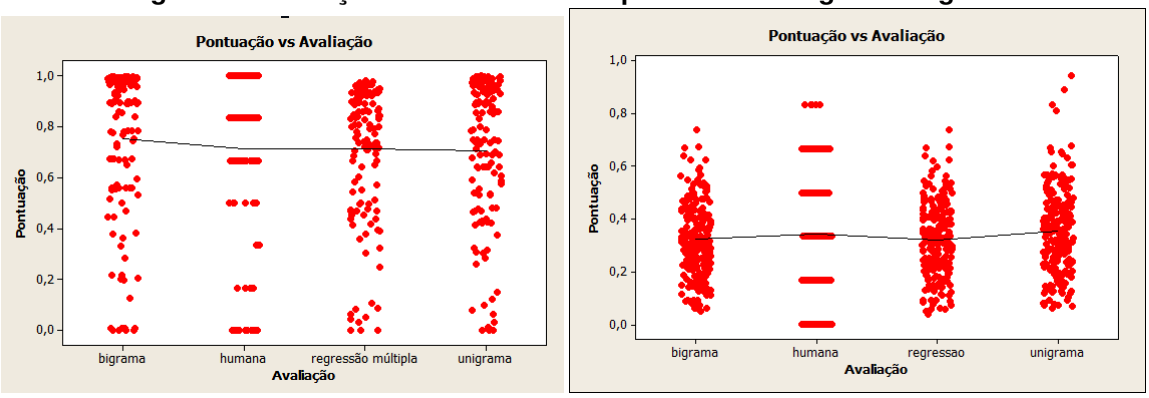
Os parâmetros que proporcionaram as melhores estimativas das pontuações para as respostas de Biologia e Geografia são dadas pela Tabela 3.

Para a disciplina de Biologia, as melhores estimativas de pontuações foram obtidas combinando unigramas e bigramas com a técnica de remoção de stop words, enquanto que para a disciplina de Geografia as melhores estimativas foram obtidas considerando todas as palavras existentes nos textos originais. Com o intuito de obter melhores resultados criamos um modelo de regressão linear múltipla para cada disciplina. No caso de Biologia, o modelo numa acurácia de 85,62%, enquanto que para Geografia resultou numa acurácia de 87,35%.

6.1. Análise de Variância

Foi feita uma Análise de Variância (ANOVA) para comparar melhores resultados obtidos por avaliações estimadas contra avaliadores humanos. A Figura 2 mostra os gráficos individuais das pontuações dos avaliadores humanos e das pontuações estimadas por unigrama e bigrama mais a regressão. Percebe-se graficamente uma pequena diferença entre as médias nas pontuações comparando-se avaliação humana e avaliações estimadas.

Figura 2. Pontuações humanas das respostas de Biologia e Geografia



Os resultados da análise de variância para o grupo de respostas de Biologia podem ser vistos na Figura 3:

Figura 3. Anova das respostas de Biologia

One-way ANOVA: Pontuação versus Avaliação

Source	DF	SS	MS	F	P
Avaliação	3	0,1760	0,0587	0,67	0,573
Error	516	45,4143	0,0880		
Total	519	45,5903			

S = 0,2967 R-Sq = 0,39% R-Sq(adj) = 0,00%

Level	N	Mean	StDev
bigrama	130	0,7527	0,2977
humana	130	0,7128	0,3340
regressão mút	130	0,7129	0,2689
unigrama	130	0,7064	0,2821

Individual 95% CIs For Mean Based on Pooled StDev

```
------+---------+---------+---------+---
      (-----------*-----------)
   (-----------*-----------)
   (-----------*-----------)
(-----------*-----------)
------+---------+---------+---------+---
    0,680     0,720     0,760     0,800
```

Pooled StDev = 0,2967

Os resultados da análise de variância para o grupo de respostas de Geografia podem ser vistos na Figura 4:

Para o grupo de respostas de Biologia tivemos $F = 0,67$ e $P = 0,573$, enquanto que pra o grupo de respostas de Geografia tivemos $F = 1,34$ e $P = 0,261$. Assim, para estas amostras analisadas concluímos que não existem diferenças significativas entre as avaliações estimadas e a avaliação humana.

Figura 4. Anova das respostas de Geografia

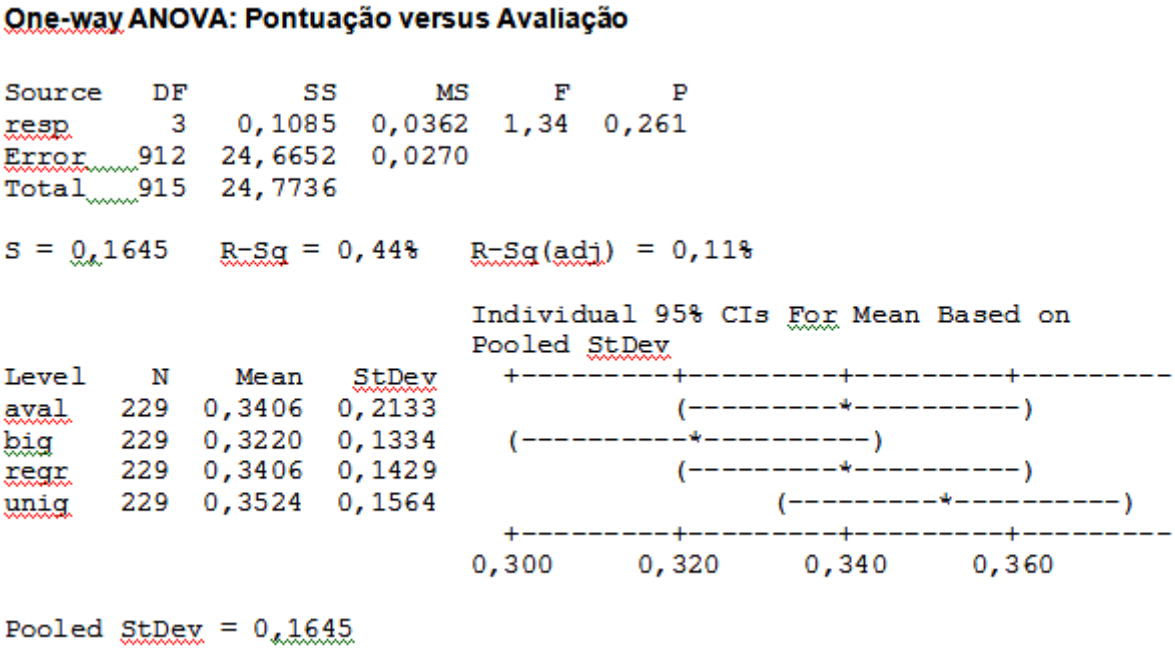


Tabela 4. Média das acurácias

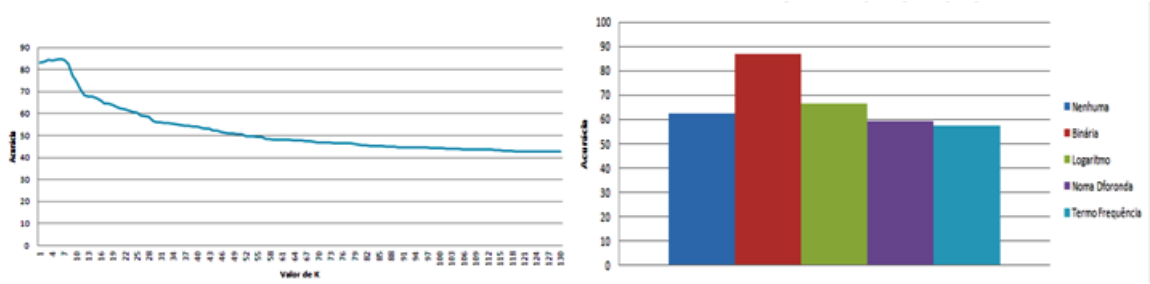
Parâmetros	Média das acurácias	
	Biologia	Geografia
Ponderação local	75,25	61,94
Ponderaçao global	84,77	82,43
Dimensão k	52,15	68,02
Medida de similaridade	84,04	83,99

6.2. Ajuste de Parâmetros

Durante a execução do método LSA proposto, em cada etapa, ajustamos sucessivamente os principais parâmetros e verificou-se a influência de cada um deles na eficácia do método. Considerando os dois grupos de respostas realizamos 89750 testes. Na Tabela 4 temos a média das acurácias nos ajustes realizados durante os testes dos parâmetros ponderação, valor da dimensão k do espaço semântico e medida de similaridade para os dois grupos de respostas:

Tomando como referência a média das acurácias obtidas durante a execução dos tetes, o parâmetro que mais influenciou a eficácia do método para as respostas de Biologia foi a dimensão k do espaço semântico, enquanto que o parâmetro ponderação local teve a maior influência para o grupo de respostas de Geografia. A ponderação global não teve nenhuma influência sobre os resultados obtidos. A Figura 5 mostra o comportamento do valor k e das ponderações locais fixados os demais parâmetros da melhor configuração. Por um lado percebe-se grandes valores de acurácias para valores baixos de k e uma piora considerável destes valores quanto maior for o valor de k; por outro lado percebe-se claramente um melhor desempenho para ponderação local binária em relação as demais ponderações locais.

Figura 5. Valores da acurácia durante o ajuste de parâmetros



7. Conclusões e trabalhos futuros

O foco deste artigo foi o ajuste de parâmetros de um método LSA para uso prático na avaliação automática de ensaios previamente avaliados por especialistas humanos comparando-se as pontuações atribuídas. Notou-se que ajuste de alguns parâmetros recomendados pela literatura consultada não melhoraram os resultados obtidos, comprovando que o uso de LSA é fortemente dependente de domínio. Uma acurácia de 87,35% e os resultados estatísticos da análise de variância habilitam LSA como um método que pode ser integrado a um Ambiente Virtual de Aprendizagem. O desenvolvimento de ferramentas que automatizem a correção de respostas discursivas, neste contexto, traz vantagens para professores e alunos: 1) feedback imediato para os alunos guiando, reorientando e estimulando seus próximos passos; 2) permite o um ranking continuo da turma, tornando o aluno ciente da sua avaliação em relação a seus colegas; 3) libera a carga de trabalho do professor com a correção manual dessas respostas; 4) permite ao professor um continuo acompanhamento da performance da turma, com a identificação de situações extremas, onde ele deverá focar seus esforços ou redirecionar o programa. O que se quer futuramente com base nestes resultados verificar melhor a dependência entre os parâmetros que foram ajustados e trabalhar com novas questões de outros domínios.

Referências

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE*, 41:391–407.

Lifchitz, A., Jhean-Larose, S., and Denhière, G. (2009). Effect of tuned parameters on an lsa multiple choice questions answering model. *Behavior Research Methods*, 41:1201–1209.

Nakov, P. (2000). Chapter 15: Getting better results with latent semantic indexing. In *In Proceedings of the Students Prenetations at ESSLLI-2000*.

Valenti, S., Neri, F., and Cucchiarelli, R. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2:2003.

Wild, F., Stahl, C., Stermsek, G., Penya, Y., and Neumann, G. (2005). Factors influencing effectiveness in automated essay scoring with lsa. In *Proceeding of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*.