Avaliação automática de respostas textuais curtas por similaridades de *n*-gramas: refinamentos por regressão linear

Silvério Sirotheau[1], João Carlos A. dos Santos, Eloi L. Favero[1]

[1]Instituto Ciências Exatas e Naturais - Universidade Federal do Pará (UFPA)
Rua Augusto Corrêa, 01 - Guamá. CEP 66075 -110 - Belém - PA – Brasil
Programa de Pós-graduação em Ciência da Computação - PPGCC

{silverio,jcas,favero}@.ufpa.br

Abstract. *In distance education, the need for intelligent virtual environments has been growing, where one of the components is a system of automatic assessment for conceptual open-ended questions. We work with answers to entrance examination questions using n-grams text-like similarity techniques and the linear regression method. The accuracy of the system was contrasted with that of the human evaluators, which resulted in 0.82 against 0.94, Biology test, and 0.86 against 0.85 Geography test. This study shows that this technology is reaching maturity to be used with great advantages in these virtual teaching environments: low cost, instant feedback, frees the teacher from the work of correction and attends large classes.*

Resumo. *No ensino a distância cresce a necessidade de ambientes virtuais inteligentes, onde um dos componentes é um sistema de avaliação automática de questões conceituais discursivas. Trabalhamos com respostas de questões do vestibular utilizando técnicas de similaridade de textos baseadas em n-gramas e o método de regressão linear. A acurácia do sistema foi contrastada com a dos avaliadores humanos, que resultou em 0.82 contra 0.94, prova Biologia, e 0.86 contra 0.85 prova Geografia. Este estudo mostra que esta tecnologia está alcançando maturidade para ser utilizadas com grandes vantagens nestes ambientes virtuais de ensino: baixo custo, feedback imediato, libera o professor do trabalho de correção e atende grandes turmas.*

1. Introdução

A avaliação automática de texto há muito tempo atrai o interesse das comunidades linguísticas, filosóficas e da teoria da informação (Hatzlvassiloglou et al. 1999). Com o avanço da tecnologia na área da educação, as instituições de ensino vem promovendo cursos abertos na modalidade à distância (EaD), tais como Coursera, Udacity, OpenClass e EdX. Nesse contexto, uma funcionalidade de avaliação automática passa a ser bem relevante (Cheniti-Belcadhi et al. 2004).

O estudo da avaliação automática de questões discursivas começou ainda na

DOI: 10.5753/cbie.sbie.2018.1433

década de 60 com o sistema PEG (Page 1966). Posteriormente surgiram outras iniciativas: E-rater (Burstein et al. 1998), IEA (Foltz et al. 1999) e Intellimetric (Learning 2000). O objetivo destes sistemas é auxiliar a avaliação humana liberando o professor da correção manual, permitindo que ele direcione sua atenção para focos mais específicos. Além disso, o custo da correção é bem baixo. Por outro lado, no contexto da *Web*, na EaD estes sistemas de avaliação automática possuem outras vantagens: i) *feedback* imediato, diminuindo o tempo de espera do aluno; ii) estar sempre disponível, permitindo repetidas submissões do estudante (Zupanc e Bosnic 2016) e; iii) permitindo também atender turmas com um grande número de alunos, pois a máquina não diminui a eficácia com uma atividade maçante e repetitiva. Nosso foco é na avaliação automática de questões conceituais discursivas do tipo resposta curta, até um parágrafo - tipo de questão que é bem útil para ser disponibilizado em ambientes virtuais (Pérez et al. 2005).

Na avaliação automática de questões discursivas existem duas principais linhas de pesquisa: a primeira é baseada em corpus e medidas de similaridade entre textos (Santos e Favero 2015), (Gomaa e Fahmy 2012), (Pribadi et al. 2017) e a segunda é baseada em métricas de similaridade entre conceitos extraídos das respostas utilizando-se técnicas de aprendizagem de máquina e processamento de linguagem natural (PLN) (Mohler e Mihalcea 2009), (Mitchell et al. 2002). Na primeira abordagem baseada na similaridade entre textos, o PLN é apenas superficial (coleta de *tokens*); enquanto que na segunda abordagem de similaridade entre redes de conceitos são necessários métodos de PLN mais sofisticados (etiquetagem, resolução de pronomes, etc).

Uma das primeiras abordagens para trabalhar com similaridade entre textos é o modelo de espaço vetorial (Salton et al. 1975) com uma ponderação do inverso da frequência/documento. Outra similar é o método da análise semântica latente - LSA (Landauer et al. 1999). Estas duas abordagens possuem bons resultados para recuperação de informação e classificação de texto (Mohler e Mihalcea 2009). A abordagem de medir a similaridade baseada em corpus de textos é independente de domínio; também é mais fácil de ser criada quando comparada com técnicas que dependem da PLN para extrair os conceitos dos textos.

O foco deste trabalho é na similaridade entre textos, buscando encontrar uma abordagem que apresente uma acurácia próxima da acurácia medida entre avaliadores humanos. Quando um sistema alcança uma acurácia próxima a dos avaliadores humanos torna-se confiável para ser utilizado na correção das questões dentro dos ambientes virtuais de ensino (Haley et al. 2007). Esforços vêm sendo feitos nesta direção (Pribadi et al. 2017).

Neste artigo apresentamos uma proposta de abordagem para avaliação automática de respostas discursivas usando técnicas de similaridade entre textos centradas em Uni(gramas) e Bi(gramas). Nesta abordagem busca-se esclarecer algumas questões. Primeiro, existe um debate sobre como criar uma boa resposta de referência

(Pérez et al. 2005). (1) Podemos pegar algumas das melhores respostas dos estudantes e considerar elas como resposta de referência? Dentre as várias técnicas de pré-processamento (Burrows et al. 2015) nós utilizamos três: de superfície (ex. remoção de pontuação), léxico (ex. correção ortográfica e remoção de *stop word*, morfológico (ex. *stemming*). (2) O pré-processamento influencia na acurácia final? Com o uso de unigramas, técnica baseada em "saco de palavras", perde-se a informação da ordem do texto nas respostas (Pérez et al. 2005). (3) Os bigramas minimizam o problema mas são poucos frequentes. Combinando unigramas com bigramas teremos uma boa acurácia?

Este artigo está organizado da seguinte forma. A seção 2 apresenta uma breve visão geral de trabalhos relacionados. A seção 3 apresenta o conjunto de dados trabalhado. A seção 4 apresenta o método proposto. A seção 5 apresenta resultados e discussão. Finalmente, a seção 6 apresenta conclusão.

2. Trabalhos Relacionados

Avaliação automática de questões discursivas é uma área de pesquisa em andamento desde a década de 1960 (Page 1966), (Hearst 2000), (Noorbehbahani e Kardan 2011). Existem vários sistemas comerciais em uso, tais como: E-rater (Burstein et al. 1998), IEA (Foltz et al. 1999) e Intellimetric (Learning 2000). Apesar destes esforços, a tecnologia não está totalmente desenvolvida a ponto de estar disponível nas plataformas virtuais de ensino, pois ainda não é uma tecnologia válida e confiável (Zupanc e Bosnic 2016). Neste contexto, pretende-se contribuir estudando como elevar a acurácia sistemas *versus* humanos (*SxH)* para valores próximos a acurácia de humano *versus* humano (*HxH)*, focando em questões curtas de até um parágrafo.

Diversos trabalhos apresentam métodos para avaliar respostas discursivas do tipo curta: Gütl (2007) construiu um sistema de avaliação para respostas curtas, fundada em uma abordagem híbrida, onde o principal componente de avaliação é baseado na comparação da resposta do estudante com uma resposta de referência utilizando o modelo espaço vetorial, no qual foram testados 368 respostas alcançando uma correlação de 0.8. Leacock e Chodorow (2003) descrevem um mecanismo de pontuação de respostas curtas a partir de resposta de referência de especialistas. A abordagem é baseada em técnicas de PLN, onde trabalhou com um corpus de 16.625 respostas e alcançou uma acurácia de 84% em relação aos humanos. Mohler e Mihalcea (2009) exploram técnicas não supervisionadas para a avaliação automática de respostas curtas. Foram combinadas medidas baseadas em conhecimento do WordNet e LSA, alcançando uma correlação de 0.50 (*SxH*) contra uma de 0.64 de (*HxH*). Rodrigues e Araújo (2012) exploraram técnicas de PLN com uma etapa de tradução de frases para formas canônicas (listas de palavras x etiqueta) via a substituição de sinônimos, como o uso de um tesauro. Na etapa de classificação utilizaram o modelo espaço vetorial e alcançaram uma correlação de 0.78 entre a média dos avaliadores e o escore dado pelo sistema. Gomaa e Fahmy (2014) utilizando diversas métricas de similaridades, algumas de

caracteres e outras de palavras, como entrada do método de classificação, numa base de 610 respostas obtiveram 0.68 (*SxH*) contra 0.86 (*HxH*) em correlação.

3. Conjunto de dados

O corpus da pesquisa foi constituído por uma coleção de respostas a duas questões discursivas que constam no edital 016/2007 do vestibular da UFPA. De um universo de mil folhas de respostas foram selecionadas as duas questões com mais folhas de respostas preenchidas: Bio(logia) com 130 respostas e Geo(grafia) com 229 respostas. O candidato escolhia um subconjunto das 26 questões que responderia, por isso não temos mil respostas por questão. A questão de Bio possui em médias de 28 palavras por resposta e a de Geo 74 palavras. Durante o processo de digitalização das respostas foram feitas apenas correções de ortografia sem alterar a concordância gramatical do texto original. Cada resposta possui a nota de dois avaliadores humanos, portanto podemos calcular a acurácia de acerto entre eles (*HxH*), ver Tabela 1.

Tabela 1. Características do conjunto de dados

Corpus	Quantidades	Palavras	$H \times H$
Bio	131	$\min = 4$, $\max = 56$, Média $= 28.48$	0.94
Geo	229	$\min = 9$, $\max = 189$, Média $= 74.56$	0.85

A Tabela 1 mostra a quantidade de respostas de cada conjunto, o mínimo de palavras, o máximo de palavras, a média de palavras e a acurácia entre avaliadores humanos.

4. Método proposto

O problema da avaliação automática baseado em técnicas de similaridade entre textos, se resume no desafio de medir a similaridade entre as respostas dos estudantes e a resposta de referência (Pribadi et al. 2017). Burrows et al. (2015) apresenta uma arquitetura para desenvolvimento de sistemas de avaliação de textos, similar a Figura 1. Estes componentes são executados em etapas de forma linear, onde a saída da etapa anterior é a entrada da etapa seguinte. O método abordado utiliza esta arquitetura.

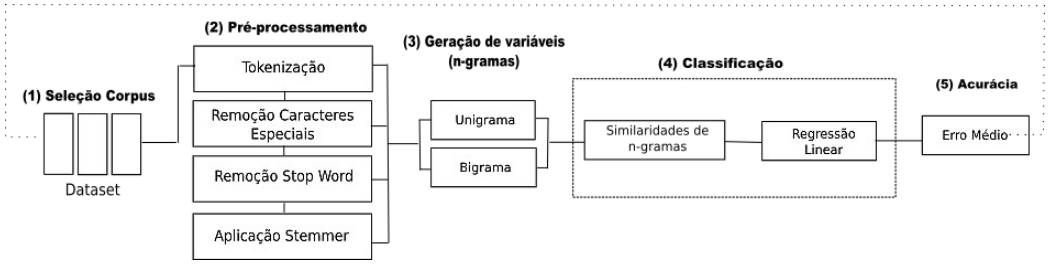

Figura 1. Arquitetura Pipeline para avaliação automática de texto.

Na etapa de pré-processamento usa-se técnicas de filtragem das respostas (transformar maiúscula em minúscula, retiradas de caracteres especiais, pontuações, mais correção de ortografia). Após a filtragem, as respostas foram "tokenizadas" em palavras para gerar dois vetores principais, um para Uni(grama) e outro para Bi(grama).

A partir destes dois vetores foram criados outros com remoção de *stop words* e com aplicação de *stemmer* (*Porter stemmer*).

Na etapa de geração de variáveis são exploradas medidas de similaridades derivadas da combinação de Uni e Bi dos *tokens* das sentenças. Foram testadas diferentes medidas de similaridade e/ou distância: Jaccard, Overlap, Dice, Cosseno, Cosseno-vetorial e Distância Euclidiana, ver Tabela 2.

Tabela 2. Medidas de similaridade de conjunto e frequência de termos

Listas de termos A e B		**Vetores numéricas A e B**	
Jaccard	$\frac{card(A \cap B)}{card(A \cup B)}$	Distância euclidiana	$\sqrt{\sum_i (a_i - b_i)^2}$
Overlap	$\frac{card(A \cap B)}{\min(card(A), card(B)}$	Cosseno-vetorial	$\frac{< A, B >}{\|A\| \cdot \|B\|}$
Dice	$\frac{card(A \cap B)}{card(A) + card(B)}$	–	
Cosseno	$\frac{card(A \cap B)}{\sqrt{card(A) \cdot card(B)}}$	–	

Combinando as variáveis são gerados vetores para a etapa de classificação, que usa regressão linear simples e múltipla para produzir o escore de cada resposta. A regressão múltipla permite combinar várias variáveis num único escore:

$$y = a_0 + a_1 x_1 + a_2 x_2 + e, \quad (1)$$

onde *y* é a variável dependente que representa o vetor de pontuação do sistema; x_1 e x_2 são variáveis independentes fornecidas pelas combinações de Uni e Bi com técnicas de pré-processamento linguístico; a_0, a_1, e a_2 são os parâmetros a serem estimados e; *e* é o erro. Na avaliação da acurácia, a saída da regressão é convertida para um valor inteiro na escala de notas atribuídas às provas, simulando o escore de um avaliador humano.

Vale ressaltar que Uni e Bi possuem características distintas e complementares. As medidas de conjunto com Uni removem as palavras repetidas, porém com Bi só são removidos os casais de palavras repetidas. Noutro aspecto, os Uni são mais frequentes entre dois textos similares, mas eles não consideram a ordem de escrita do texto e são chamados “sacos de palavras”; já os Bi consideram a ordem de escrita do texto, mas são menos frequentes.

Tabela 3. Exemplo de resposta de referência (RR) e de resposta de um estudante (RE) com Uni(grama) e Bi(grama).

Resposta	**Texto**	***Card***
RR_1 Uni	[Tecido, muscular, esqueletico, responsavel, sustentacao, epitelial, movimentacao, corpo, Adiposo, Absorcao, impactos, mecanicos, osseo, protecao, externa, corpo, controle, temperatura]	18
RR_2 Bi	[(Tecido, muscular),(muscular, esqueletico), (esqueletico, responsavel),(responsavel, sustentacao),(sustentacao, epitelial)(epitelial, movimentacao),(movimentacao, corpo),(corpo, Adiposo),(Adiposo, Absorcao),(Absorcao, impactos),(impactos, mecanicos),(mecanicos, osseo),(osseo, protecao),(protecao, externa), (externa, corpo),(corpo, controle),(controle, temperatura)]	17
RE_1 Uni	[tecido, epitelial, responsavel, protecao, externa, pele, osseo]	7
RE_2 Bi	[(tecido, epitelial),(epitelial, responsavel),(responsavel, protecao), (protecao, externa), (externa, pele), (pele, osseo)]	6

A Tabela 3 ilustra um exemplo (só uma parte do texto é mostrada) do conjunto de dados de Bio utilizada em nossos experimentos para demonstrar similaridade entre uma resposta de estudante (RE) e a resposta de referência (RR). Considerando os *n*-gramas da tabela 3, Dice de Uni na fórmula $| RR_1 \cap RE_1 | / | RR_1 | + | RE_1 |$ resulta em 6/(18+7)=0.24; Jaccard de Bi na fórmula $| RR_2 \cap RE_2 | / | RR_2 \cup RE_2 |$ resulta em 1/22=0.04. Na coleta de métrica os textos são substituídos por vetores numéricos que são entradas para os modelos de regressão. Para apresentação dos resultados foram selecionadas as combinações de técnicas que geraram a melhor acurácia.

5. Resultados e discussão

A meta é maximizar a acurácia *SxH* buscando uma aproximação com a acurácia *HxH*. Burrows et al. (2015) apresentam duas medidas para avaliação da acurácia de dados contínuos: correlação e erro médio. Avalia-se acurácia do método utilizando o erro médio:

$$acuracia = \frac{6 - erro_{medio}}{6} \cdot 100. \quad (2)$$

A Figura 2 apresenta as pontuações para a questão de Bio, para uma amostra aleatória de 30 respostas contra o método proposto, considerando Uni+Bi alcançou 80.43 contra acurácia entre os avaliadores humanos de 94.0.

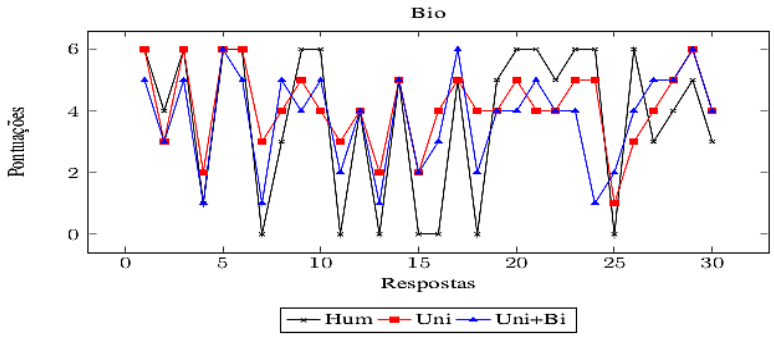

Figura 2. Uni(grama) vs Uni(grama)+Bi(grama) vs Hum(ano)

A combinação de Uni+Bi teve o melhor resultado. O uso de Bi é interessante porque foge do problema de "saco de palavras" o que torna o método robusto, pois considera a ordem de escrita do texto. Da Figura 2 pode-se observar que os gráficos de Uni e Uni+Bi tem um comportamento similar ao gráfico de Hum(ano), tendo muitas coincidências.

A Figura 3 apresenta as pontuações para a questão de Geo, para uma amostra aleatória de 30 respostas contra o método proposto considerando Uni alcançou uma acurácia de 86.08 e Uni+Bi que alcançou 86.25. A acurácia entre os avaliadores humanos foi de 85.0. Neste caso o sistema superou os avaliadores humanos. Os

resultados da Figura 2 e Figura 3 foram obtidos com remoção de *stop word* e aplicação de *stemming*, como de fato existe na literatura um consenso de que melhores resultados em processamento de textos são obtidos com remoção de *stop words* e *stemming* (Nakov et al. 2004).

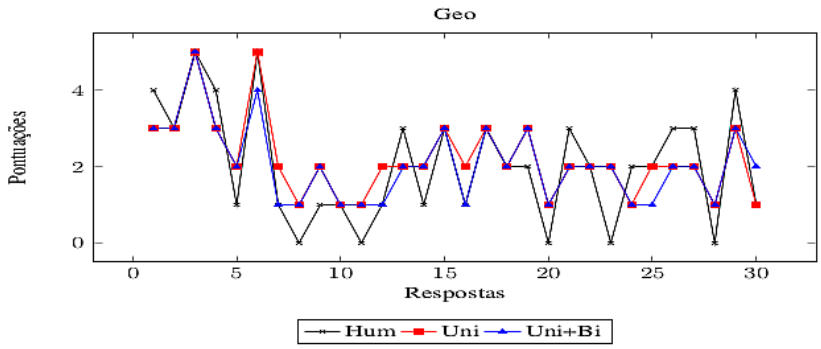

Figura 3. Uni(grama) vs Uni(grama)+Bi(grama) vs Hum(ano)

Em relação as questões levantadas, Burrows et al. (2015) relatam que a pontuação automática depende da qualidade da resposta de referência. Para as respostas de Bio existia uma resposta de referência dada por um especialista humano. Podemos compor uma resposta de referência a partir das melhores respostas avaliadas dentro do corpus? Para responder essa questão foram feitos experimentos juntando as quatro melhores respostas que resultaram numa acurácia de 0.82 contra 0.84 da resposta do especialista, portanto pelos dados podemos criar a resposta de referência a partir das melhores respostas do corpus.

Na etapa de pré-processamento foram utilizadas três técnicas de processamento morfológico: (1) remoção de Caracteres Especiais e Pontuação (+RCE); (2) remoção de *stop words* (+RSW); e (3) remoção de sufixos (*stemming*) (+RSU). Estas três técnicas foram combinadas de quatro formas: i) sem pré-processamento (-RCE, -RSW, -RSU), com remoção de caracteres especiais (+RCE, -RSW, -RSU), com remoção de caracteres especiais e *stop word* (+RCE, +RSW, -RSU) e com remoção de caracteres especiais, *stop word* e aplicação de *stemmer* (+RCE, +RSW, +RSU). Na Figura 4 temos os resultados obtidos para Bio e Geo considerando as variações nas técnicas de pré-processamento.

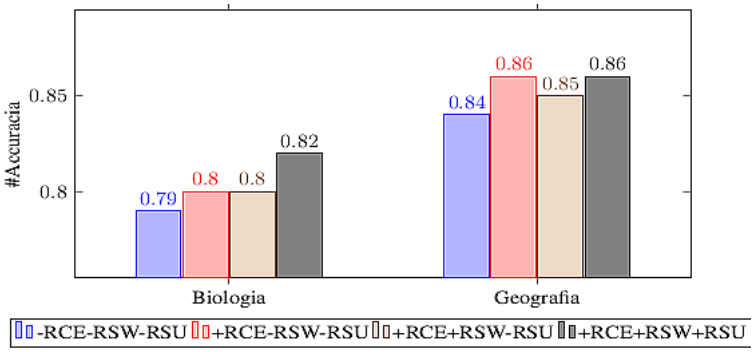

Figura 4. Compara as técnicas de pré-processamento

As diferentes técnicas de pré-processamento apresentam diferentes valores de acurácia. No entanto, as diferenças não são tão significativas, sendo a diferença do menor para o maior valor 0.03 para Bio e 0.02 para Geo. Na coleta das métricas utilizamos medidas de similaridade de teoria dos conjuntos e também de frequência de termos. As duas melhores medidas baseadas em teoria dos conjuntos foram Dice e Jaccard, tendo resultados similares aos das medidas de frequência de termos, ver Figura 5. As medidas de frequência dos termos sempre obtiveram os valores máximos.

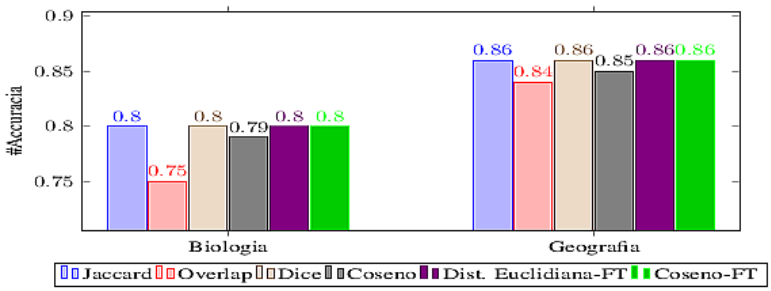

Figura 5. Comparativo das métricas

A Tabela 4 mostra um comparativo dos resultados para Uni, Bi e a combinação de Uni+Bi. Verifica-se que o uso só de Bi dá um resultado de acurácia final um pouco inferior para as duas provas; considerando esta diferença devemos ponderar o uso de Bi, pois eles fogem do problema de "saco de palavras". Por outro lado, tivemos um resultado um pouco melhor com o uso combinado de Uni+Bi: para Bio a melhor acurácia foi 0.82 e para Geo foi de 0.86; pelos dados além de manter a acurácia, esta combinação permite que a solução seja mais robusta, pois considera a ordem de escrita dos textos.

Tabela 4. Comparativo $S \times H$ vs $S \times H$

Corpus	$S \times H$ **- Uni**	$S \times H$ **- Bi**	$S \times H$ **- Uni+Bi**	$H \times H$
Bio	0.80	0.77	0.82	0.94
Geo	0.85	0.84	0.86	0.85

A Tabela 4 também mostra a coluna da acurácia entre especialistas humanos, *HxH*. No comparativo com a acurácia sistema *vs* humano *SxH*, para a prova de Bio ficou um pouco longe com 0.82 *SxH*, contra 0.94 *HxH*; porém para a prova de Geo obteve-se o resultado de 0.86 *SxH* contra 0.85 *HxH*. Este segundo resultado, o sistema superou a acurácia entre dois humanos; medindo a acurácia *SxH*, o *H* é a média dos dois humanos e medindo a acurácia *HxH*, cada *H* é a nota do especialista humano.

6. Conclusão

O objetivo deste trabalho foi testar uma proposta de avaliação automática de respostas discursivas curtas baseadas na similaridade entre textos. Combinamos Unigramas e Bigramas por regressão linear múltipla e observamos que esta combinação produziu bons resultados para respostas curtas. Os experimentos produziram acurácias *SxH* 0.82

contra *HxH* 0.94 na prova de Bio e acurácia *SxH* 0.86 contra *HxH* 0.85 na prova de Geo. Estes resultados mostram o potencial desta tecnologia para uso prático em ambientes virtuais de aprendizagem. Como trabalhos futuros temos duas frentes: estamos testando o método dentro de um ambiente virtual real e iniciamos um estudo de avaliação automática das redações do vestibular.

7. Referências

Burrows, S.; Gurevych, I.; Stein, B. (2015) The eras and trends of automatic short answer grading. International Journal of Artificial Intelligence in Education, v. 25, n. 1, p. 60-117.

Burstein, J. (1998) et al. Automated scoring using a hybrid feature identification technique. In: Proceedings of the 17th international conference on Computational linguistics-Volume 1. Association for Computational Linguistics, p. 206-210.

Cheniti-Belcadhi, L. et al. (2004) A Generic Framework for Assessment in Adaptive Educational Hypermedia. In: ICWI. p. 397-404.

Santos, J. C.; Favero, E. L. (2015) Practical use of a latent semantic analysis (LSA) model for automatic evaluation of written answers. Journal of the Brazilian Computer Society, v. 21, n. 1, p. 21.

Foltz, P. W., Laham, D. e Landauer, T. K. (1999). The intelligent essay assessor: Applications to educational technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, v. 1, n. 2, p. 939-944.

Gomaa, W. H. e Fahmy, A. A. (2012). Short answer grading using string similarity and corpus-based similarity. International Journal of Advanced Computer Science and Applications (IJACSA), v. 3, n. 11.

Gomaa, W. H. e Fahmy, A. A. (2014). Automatic scoring for answers to Arabic test questions. Computer Speech & Language, v. 28, n. 4, p. 833-857.

Gült, C. (2007). e-Examiner: towards a fully-automatic knowledge assessment tool applicable in adaptive e-learning systems. In: Proceedings of the 2nd international conference on interactive mobile and computer aided learning. p. 1-10.

Haley, D. T. et al. (2007)Seeing the whole picture: evaluating automated assessment systems. Innovation in Teaching and Learning in Information and Computer Sciences, v. 6, n. 4, p. 203-224.

Hatzlvassiloglou, V., Klavans, J. L. e Eskin, E. (1999). Detecting text similarity over short passages: Exploring linguistic feature combinations via machine learning. In: 1999 Joint SIGDAT conference on empirical methods in natural language processing and very large corpora.

Hearst, M. A. (2000). The debate on automated essay grading. IEEE Intelligent Systems

and their Applications, v. 15, n. 5, p. 22-37.

Landauer, T. K., Foltz, P. W. e Laham, D. (1998). An introduction to latent semantic analysis. Discourse processes, v. 25, n. 2-3, p. 259-284.

Leacock, C. e Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. Computers and the Humanities, v. 37, n. 4, p. 389-405.

Learning, V. (2000). A study of expert scoring and IntelliMetric scoring accuracy for dimensional scoring of Grade 11 student writing responses (RB-397). Newtown, PA: Vantage Learning.

Mitchell, T., Russell, T., Broomhead, P. e Aldridge, N. (2002). Towards robust computerised marking of free-text responses.

Mohler, M. e Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pages 567–575. Association for Computational Linguistics.

Nakov, P., Valchanova, E. e Angekova, G. (2004). Towards deeper understanding of the latent semantic analysis performance. Amsterdam studies in the theory and history of linguistic science, p. 297.

Noorbehbahani, F. e Kardan, A. (2011). The automatic assessment of free text answers using a modified BLEU algorithm. Computers & Education, v. 56, n. 2, p. 337-345.

Page, E. B. (1966). The imminence of grading eessay by computer. The Phi Delta Kappan.

Pérez, D. et al. (2005). About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment. Revista signos, v. 38, n. 59, p. 325-343.

Pribadi, F. S. et al. (2017). Automatic short answer scoring using words overlapping methods. In: AIP Conference Proceedings. AIP Publishing. p. 020042.

Rodrigues, F. e Araújo, L. (2012) Automatic Assessment of Short Free Text Answers. In: CSEDU (2). p. 50-57..

Salton, G., Wong, A. e Yang, C. (1975). A vector space model for automatic indexing. Communications of the ACM, v. 18, n. 11, p. 613-620.

Zupanc, K. e Bosnic, Z. (2016). Advances in the field of automated essay evaluation. Informatica, v. 39, n. 4.