Avaliação Automática de Ensaios, em português, centrada em atributos linguísticos de superfície e de conteúdo

Silvério Sirotheau[1,3], **João Carlos A. dos Santos**[2,3], **Eloi L. Favero**[1,3]

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

[2]Faculdade de Matemática – UFPA

[3]Laboratório de Inovação Interdisciplinar – LabX

{silverio,jcas,favero}@ufpa.br

Abstract. *There is growing need for intelligent environments for distance learning. One of its elements is a system of automatic evaluation of conceptual discursive issues. In this work, we propose a method of automatic evaluation of a test in the Portuguese language based on the refinement of content, coherence and surface statistics features to predict the score of an essay. The accuracy of the system was contrasted with the accuracy measured between two human evaluators (HxH), which resulted in an average error value of 0.91 SxH versus 0.89 HxH and a quadratic kappa accuracy of 0.62 SxH versus 0.52 HxH.This study shows that this technology is reaching maturity for use in environments.*

Resumo. *Cresce a necessidade de ambientes inteligentes para o ensino a distância. Um dos seus elementos é um sistema de avaliação automática de questões conceituais discursivas. Neste trabalho, propõe-se um método de avaliação automática de ensaio na língua portuguesa baseado no refinamento de atributos de conteúdo (semânticos), de coerência e estatísticos de superfície para predizer a pontuação de um ensaio. A acurácia do sistema (SxH) foi contrastada com a acurácia medida entre dois avaliadores humanos (HxH), o que resultou num valor de erro médio de 0.91 SxH contra 0.89 HxH e numa acurácia kappa quadrático de 0.62 SxH contra 0.52 HxH. Este estudo mostra que esta tecnologia está alcançando maturidade para o uso em ambientes.*

1. Introdução

O avanço da tecnologia na área da educação vem promovendo em instituições de ensino a modalidade de cursos abertos em plataformas *online*, tais como *Coursera*, *Udacity*, *OpenClass* e *edX*. A aplicação de avaliações compostas por textos discursivos tem grande relevância nessas plataformas pois verificam resultados de aprendizagem do aluno, e em particular seu desempenho na escrita [Zupanc and Bosnic 2016, Page 1966]. Neste contexto, uma funcionalidade de avaliação automática passa a ser bem relevante [Cheniti-Belcadhi et al. 2004].

[Shermis and Burstein 2003] definem avaliação automática como um processo de pontuação e avaliação por meio de programas de computador. [Miller et al. 2013] definem como um problema de previsão, que avalia e pontua automaticamente as soluções

fornecidas pelos alunos, comparando-as com uma solução de referência por meio de programas de computador. O objetivo destes programas é auxiliar a avaliação humana liberando o professor da correção manual, permitindo que ele direcione sua atenção para focos mais específicos do processo ensino aprendizagem. No caso da avaliação automática, existem várias linhas de pesquisas em textos discursivos como: *Automatic Short Answer Grading* (ASAG), *Automated Writing Evaluation* (AWE) e *Automated Essay Evaluation* (AEE). Neste trabalho, o ponto central é avaliação de ensaios (AEE).

[Zupanc and Bosnic 2017] definem ensaios como composições literárias curtas sobre um determinado assunto. [Valenti et al. 2003] considera ensaios como uma ferramenta útil para avaliar os resultados da aprendizagem, dando oportunidade aos alunos de mostrar suas habilidades e conhecimentos de escrita.

Muitas abordagens tradicionais no campo AEE, como *E-rater* [Burstein et al. 1998], *Intelligent Essay Assessor* (IEA) [Foltz et al. 2013] e *IntelliMetric* [Schultz 2013] diferem em seus métodos de avaliação principalmente na quantidade e tipos de atributos (*features*) coletados. Uma crítica a estas abordagens é que grande maioria dos atributos são apenas superfície linguística, i.e., não consideram a semântica do conteúdo do texto, nem a coerência do discurso [Palma and Atkinson 2018].

Neste trabalho buscou-se uma abordagem que apresente acurácia próxima da acurácia medida entre dois avaliadores humanos (HxH) por meio de um método que baseia-se na combinação de atributos linguísticos de superfície, com os semânticos e de coerência, pois quando um sistema alcança uma acurácia contra humanos (SxH) próxima a medida entre dois avaliadores humanos (HxH) torna-se confiável para ser utilizado na correção das questões, dentro dos ambientes virtuais de ensino [Haley et al. 2007].

Este artigo está organizado da seguinte forma: A seção 2 apresenta os trabalhos relacionados. A seção 3 apresenta as questões de pesquisa. A sessão 4 apresenta a metodologia. A seção 5 apresenta os resultados e a discussão. Finalmente, a seção 6 apresenta a conclusão e trabalhos futuros.

2. Trabalhos relacionados

A pesquisa de AEE iniciou-se a partir década de 1960; um dos primeiros trabalhos nesta área desenvolveu o sistema *Project Essay Grade* (PEG) com foco em avaliar as habilidades do estilo de escrita [Page 1966]. Porém, somente a partir dos anos 90, com o uso de técnicas de Processamento de Linguagem Natural (PLN) houve um avanço considerável nesta área. Apesar destes esforços, a tecnologia não está totalmente desenvolvida a ponto de estar disponível nas plataformas virtuais de ensino, pois ela ainda não é confiável. Neste contexto, pretende-se contribuir para elevar a acurácia SxH a valores próximos a HxH em respostas do tipo ensaio.

Diversas são as abordagens para avaliar respostas discursivas do tipo ensaio: O sistema *E-rater* identifica e extrai 10 classes de atributos [Attali and Burstein 2006, Burstein et al. 2004]; usa-se um modelo de regressão para atribuir uma pontuação final nos ensaios; O *Intelligent Essay Assessor* (IEA) [Laham et al. 2000] baseia-se num modelo de previsão fundado em Análise Semântica Latente (LSA); usam-se técnicas de aprendizagem de máquina para extrair significados de palavras e de documentos por meio da análise de grandes corpora de texto; requer um treinamento com uma amostra representativa (entre 200 e 500) de ensaios com pontuação humana. O *IntelliMetric* [Schultz 2013]

analisa mais de 400 atributos, parte sobre o conteúdo e parte sobre a estrutura; usam-se vários modelos matemáticos para prever a pontuação final: análise linear, abordagem Bayesiana e LSA; precisa ser treinado em pelo menos 300 ensaios avaliados por humanos. O sistema *Bookette* [Rich et al. 2013] usa PLN para obter cerca de 90 atributos; aplicam-se redes neurais como modelo de predição; é treinado num conjunto de 250 a 500 ensaios com pontuação humana.

Trabalhos contemporâneos relatam métodos de avaliação automática que focam em três dimensões de atributos linguísticos: estatísticos de superfície, de conteúdo semântico e de coerência e consistência. [Zupanc and Bosnic 2017] propõem uma extensão do sistema *Semantic Automated Grader for Essays* - SAGE (baseado em conteúdo e atributos de superfície) incorporando novos atributos para medir coerência e consistência (101 atributos); usam-se técnicas de regressão de árvores de decisão para predição do escore; [Palma and Atkinson 2018] propõem um método baseado na coerência do discurso mesclando modelos semânticos e sintáticos com extração de 54 atributos; eles também usam técnicas de regressão de árvores de decisão.

Apesar de uma ampla variedade de sistemas, não há muitos trabalhos publicados sobre quais são as contribuições de cada atributo para a precisão de um sistema de avaliação [Vajjala 2018]. Além disso, poucos estudos têm sido publicados sobre AEE para ensaios em Português. Neste contexto, pretendemos contribuir para elevar a acurácia SxH para valores próximos a HxH, focando no refinamento de atributos para ensaios na língua portuguesa.

3. Questões de pesquisa

Durante a elaboração de um levantamento bibliográfico para analisar os resultados dos trabalhos de AEE verificam-se que pouco se sabe sobre quais atributos são bons preditores da qualidade do ensaio:

Q1) Publicações voltadas para ensaios [Zupanc and Bosnic 2017, Palma and Atkinson 2018, Vajjala 2018] em língua inglesa, relatam atributos compreendendo três classes (de estatísticos de superfície, de conteúdo e de coerência); será possível portar as técnicas desses estudos para a língua portuguesa?

Q2) Quais atributos contribuem mais para a precisão de um sistema de avaliação automática para ensaios em português?

Q3) A importância de contribuição dos atributos repete-se em diferentes conjuntos de dados?

4. Metodologia

Os sistemas mais modernos em AEE combinam vários recursos (seleção de corpus, pré-processamento, extração de atributos, modelo de predição, validação, verificação da acurácia) para construção de uma metodologia, tal como a da arquitetura *pipeline* com 6 etapas da Figura 1.

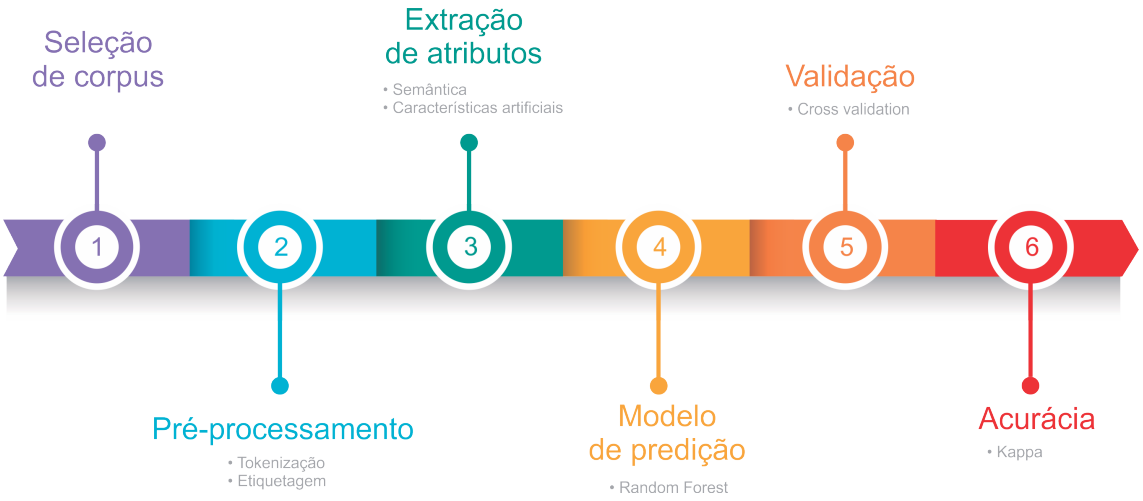

Figura 1. *Pipeline* da arquitetura do sistema utilizado.

4.1 - Seleção de corpus

Na etapa 1 utilizam-se dois conjuntos de dados. O primeiro conjunto de dados vem da competição *Automated Essay Scoring* (AES) da *Kaggle* (http://kaggle.com) que fornece uma variedade de tipos de ensaios. Este corpus possui oito corpora escritos por estudantes do ensino médio (7º ao 10º ano). Cada ensaio foi avaliado por dois especialistas humanos, atribuindo-se um *score* final; exceto o conjunto de ensaios n. 2, que foram atribuídos dois *scores* finais - dois critérios (escrita e habilidades de linguagem). Cada um dos oito corpora têm suas próprias características, como mostrado na tabela 1.

Tabela 1. Detalhes do conjunto de dados de língua Inglesa

Ensaio	Tipo	Quant.	Pontuação	Média palavras	Média notas
1	Persuasivo	1783	2-12	366.40	8.53
2*	Persuasivo	1800	1-6,1-4	381.19	3.42-3.33
3	Baseado em fonte	1726	0-3	108.69	1.85
4	Baseado em fonte	1772	0-3	94.39	1.43
5	Baseado em fonte	1805	0-4	122.29	2.41
6	Baseado em fonte	1800	0-4	153.64	2.72
7	Expositivo	1569	2-24	171.28	16.06
8	Narrativo	723	10-60	622.13	36.95

Muitos conjuntos de dados dos sistemas de AEE, como o da Tabela 1, possuem dois *scores* para cada ensaio, dados por avaliadores humanos independentes; neste, pode-se medir a acurácia entre eles para ser contrastada com a acurácia obtida pelo sistema.

O segundo conjunto de dados é uma coleção de ensaios do Concurso Público para cargos de carreira de técnico administrativo em educação da Universidade Federal do Oeste do Pará (UFOPA) sob o tema a ser desenvolvido "A atual crise político-social do Brasil e ações políticas para seu enfrentamento". A tabela 2 apresenta algumas características do corpus.

Tabela 2. Descrição do conjunto de dados de ensaios

Tipo	Número de Ensaios	Pontuação	Média de notas	Média palavras
Dissertativa	859	0-10	6.0	45.60

Na figura 2, apresentamos a visualização das 40 palavras mais frequentes na base de dados de ensaio na língua portuguesa, no total de 1.323.268 palavras.

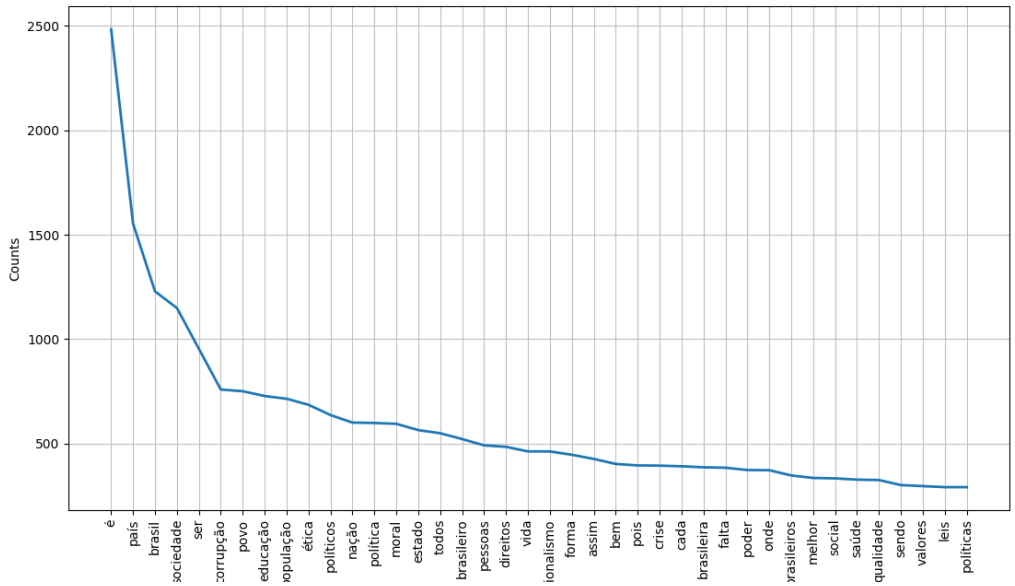

Figura 2. Palavras mais frequentes na base de dados de português

4.2 - Pré-processamento

Na etapa de pré-processamento, cada ensaio foi segmentado em sentenças e tokenizadas usando a biblioteca *Natural Language Toolkit* (NLTK)[Perkins 2014]. As *stop words* foram removidas para facilitar a extração de informações, assim como os *tokens* foram etiquetados conforme suas categoria gramaticais (tanto no inglês como no português).

4.3 - Extração de atributos

Nos sistemas de AEE existe uma variedade de atributos que são utilizados para descrever características que medem a qualidade do texto. Para formar os conjuntos de dados para o aprendizado supervisionado foram extraídos os seguintes atributos:

- **Atributos de Conteúdo**: São extraídos por meio de um modelo de Análise Semântica Latente (LSA) que faz decomposição de matrizes e projeção de vetores, criando um espaço semântico, para comparação textual; e n-gramas que geram vetores semânticos com a sequência de palavras a partir dos conjuntos de dados do ensaio, os quais também medem a similaridade dos textos.

- **Atributos estatísticos de superfície:** São atributos extraídos de um texto usando métricas estatísticas, como frequência de palavras, comprimento de palavras, frequência de etiquetas sintáticas, etc.

- **Atributos de Coerência:** Atributos de coerência descrevem o fluxo de informações de uma parte do ensaio para outra. Existe uma abordagem tradicional para modelar coerência do texto conhecida como Teoria do discurso, que pode ser realizada por três modelos: Teoria do Centro [Grosz et al. 1995], Teoria da Estrutura Retórica [Mann and Thompson 1988, Matthiessen and Thompson 1988] e Modelos Semânticos [Carel and Ducrot 2001]. Neste trabalho utilizaram-se quatro métodos (Figura 3) baseados na Teoria do Centro, levando em consideração a ideia de janela [Zupanc and Bosnic 2017, Palma and Atkinson 2018]: um ensaio é decomposto em várias janelas; com ou sem sobreposição; janelas são contrastadas entre si para se avaliar a coerência.

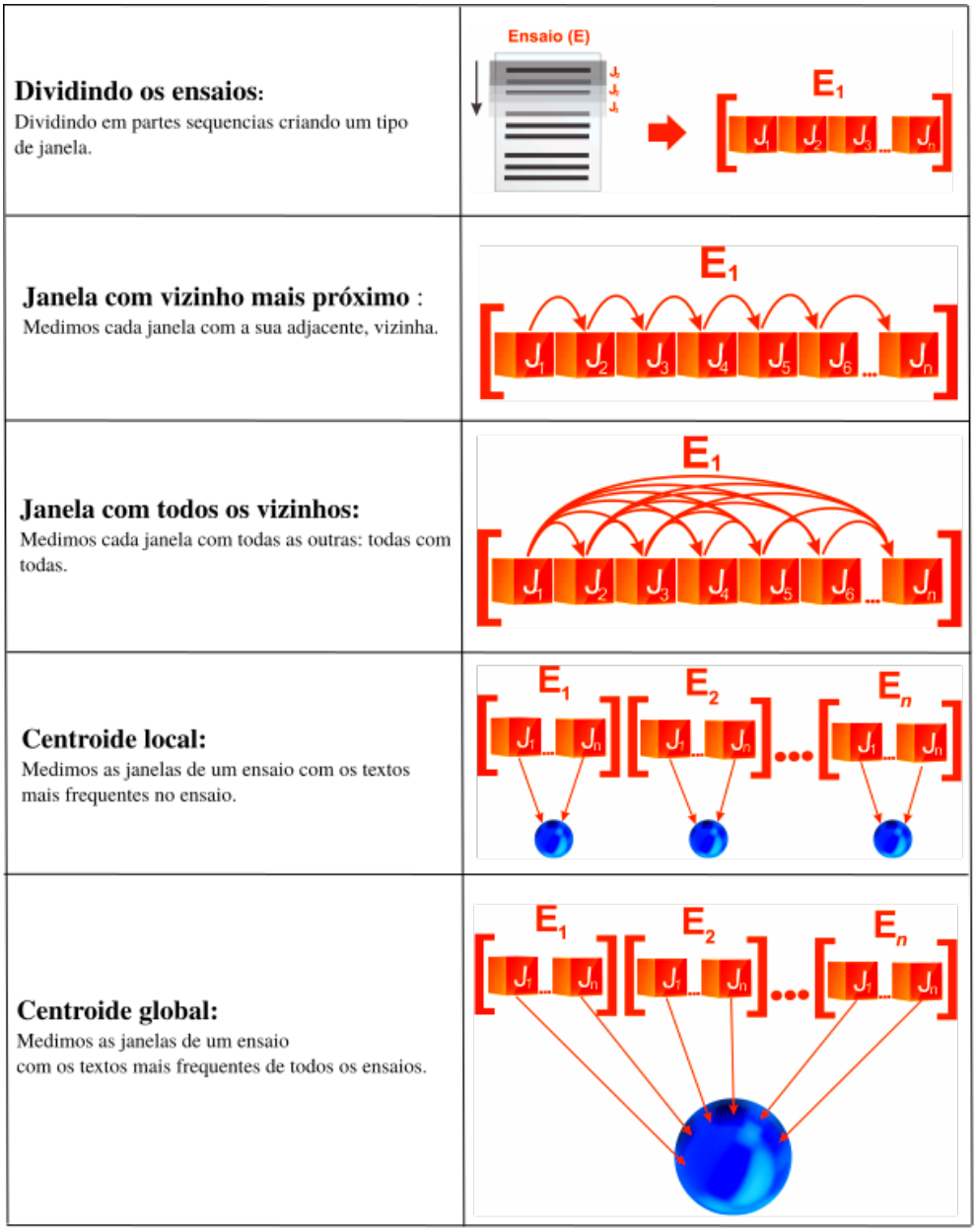

Figura 3. Esquema de quatro métodos para a Teoria do Centro.

Os métodos da Figura 3 são avaliados com LSA (unigramas) e com n-gramas

(bi-gramas). O LSA utiliza unigramas onde a ordem das palavras não é considerada na representação espacial; mas nos n-gramas, os bi-gramas consideraram também a ordem das palavras no texto; como as janelas de sequencias de palavras possuem sobreposição os bi-gramas possuem uma frequência mínima. Por outro lado, testaram-se também duas abordagens para definição de janelas: sequência de palavras e sequência de sentenças.

4.4 – Modelo de predição

Na etapa seguinte, o algoritmo de *Random Forest* é utilizado para construir um modelo de predição. Um novo ensaio (não visto no treinamento) é então avaliado automaticamente a partir da coleta dos seus atributos.

Para trabalhar com grandes dimensionalidades de atributos o *Random Forest* tem se destacado [Zupanc and Bosnic 2017, Palma and Atkinson 2018]; ele possui as seguintes vantagens:

1. poder ser usado tanto em regressão quanto em classificação;
2. trabalha com uma grande quantidade de atributos, podendo identificar os mais significativos;
3. estima com eficiência os dados ausentes mantendo a precisão e;
4. possui vários métodos que podem ser usados para balancear erros no conjunto de dados.

Para coleta da acurácia rodamos o modelo com a técnica de *Cross Validation*, com 10 *folds*. A acurácia coletada é a média dos 10 testes.

4.5 – Métricas de avaliação

Para validar o modelo utilizou-se Kappa Ponderado Quadrático (KPQ) [Fleiss and Cohen 1973], que mede o grau de concordância entre duas classes e é análogo a um coeficiente de correlação. Essa métrica geralmente varia de 0 (pouca concordância entre avaliadores) a 1 (concordância completa entre avaliadores). Caso a concordância entre os avaliadores seja abaixo do mínimo esperado, essa métrica também pode resultar em valores negativos.

O KPQ é calculado criando-se uma matriz de acordo com a equações 1 e 2. Neste caso, cada célula da matriz O, i.e. $O_{i,j}$, corresponde a um ensaio pontuado i do avaliador humano e j do sistema. $W_{i,j}$ contém os pesos calculados conforme a Equação 1 e a matriz E contém as pontuações esperadas, i.e., dos avaliadores humanos, obtidas pela multiplicação dos vetores de histograma das duas pontuações.

$$W_{i,j} = \frac{(i-j)^2}{(N-1)^2} \tag{1}$$

No final do processo o Kappa é calculado como:

$$k = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}} \tag{2}$$

Outra métrica utilizada para validar o modelo é o erro médio que de acordo com a equação 3 mede a diferença entre a pontuação atribuída pelo sistema e a pontuação do avaliador humano, onde $e_i = |x_i - y_i|$, sendo x_i pontuação real e y_i a pontuação estimada para cada resposta i, com $i = 1, ..n$, total de respostas do conjunto de teste.

$$acc = 1 - \frac{\sum_{i=1}^{n} e1}{n} \tag{3}$$

O resultado do experimento de cada modelo é validado contra a acurácia dos avaliadores humanos (HxH). Ajustes, na coleta de atributos e nos parâmetros, são feitos buscando-se que a acurácia SxH supere a acurácia HxH.

5. Resultados e discussão

Em um primeiro momento, aplicou-se essa abordagem nos ensaios da língua inglesa. Neste corpus, ajustaram-se os nossos modelos contrastando com os resultados da literatura, buscando se possível superar a acurácia . Os resultados são consolidados na Tabela 3.

Tabela 3. Resultados do conjunto de dados de ensaios na língua inglesa.

	1	**2a**	**2b**	**3**	**4**	**5**	**6**	**7**	**8**
Kappa (SxH)	0.84	0.69	0.60	0.67	0.72	0.79	0.71	0.78	0.70
Kappa (HxH)	0.72	0.81	0.80	0.77	0.85	0.75	0.78	0.72	0.62
Erro (SxH)	0.93	0.96	0.93	0.91	0.92	0.89	0.90	0.89	0.92
Erro (HxH)	0.91	0.91	0.80	0.79	0.74	0.80	0.83	0.83	0.91

Em relação à concordância KPQ obtiveram-se os melhores desempenhos em quatro (1,5,7,8) dos nove conjuntos de dados; nos outros obteve-se um desempenho próximo. Por outro lado, em relação ao erro médio, obtiveram-se melhores resultados em todos os nove conjuntos de dados.

Em um segundo momento, utilizou-se o modelo já ajustado para o corpus de ensaios em português. Para isso, foram feitos ajustes de adaptação da língua: etiquetagem, *stop words*, *stemmer*, etc. A tabela 4 consolida os resultados, onde a acurácia sistema *vs* humano (SxH) supera com uma boa diferença a acurácia entre os dois avaliadores humanos (HxH), tanto para KPQ como para erro médio.

Tabela 4. Resultados do conjunto de dados de ensaios na língua portuguesa.

Medidas	**Ensaio**
Kappa (SxH)	0.62
Kappa (HxH)	0.52
Erro (SxH)	0.91
Erro (HxH)	0.89

Em relação às questões de pesquisa levantadas, Q1) Publicações voltadas para ensaios [Zupanc and Bosnic 2017, Palma and Atkinson 2018, Vajjala 2018] em língua inglesa, relatam atributos compreendendo três classes (de estatísticos de superfície, de

conteúdo e de coerência); será possível portar as técnicas desses estudos para a língua portuguesa? Para checar à portabilidade do método, realizamos experimentos com a abordagem aqui proposto em dois conjuntos de dados. O primeiro com 12798 ensaios na língua inglesa e o segundo com 859 ensaios na língua portuguesa. Os ensaios foram categorizados em conjuntos com *scores* de intervalos diversos conforme tabela 1 e 2. Os ensaios tinham dois *scores* humanos, então mediu-se a acurácia entre os dois humanos e, uma vez rodados os experimentos com o método proposto, medimos a acurácia do sistema contra a média dos dois humanos.

Os resultados do experimento permitem ver que em o sistema produziu resultados com acurácia 0.10 acima dos avaliadores humanos, mostrando que a ideia de portabilidade do método é bem promissora, o que responde à questão de pesquisa Q1.

(Q2) Quais atributos contribuem mais para a precisão dos AEE para o português? Para responder essa questão realizou-se um experimento com 72 atributos, resultando nas 28 principais atributos da Tabela 5 ordenadas por importância.

Tabela 5. Resultados do conjunto de dados de ensaios na língua portuguesa.

Atributo	Importância	Atributo	Importância
Total de palavras longas	0.14	Total de sentenças longas	0.02
Total de caracteres	0.13	Total de sentenças curtas	0.02
Total de silabas	0.06	Média do total de sentenças	0.02
Total de Adjetivos	0.06	Média de Freq. de palavras	0.02
Densidade Lexical	0.05	Total de Nominais	0.02
Hapax	0.05	Total de ADP	0.02
Cosseno Vet. Unigrama	0.04	Total de Adverbio	0.02
Total de palavras curtas	0.03	Total de Conjunções	0.02
Total de DET	0.03	Total de Pronomes	0.02
Dist. Euclidiana Unigrama	0.03	Knn melhores *scores*	0.02
Total de palavras	0.02	Total de Numerais	0.01
Média do tam. das palavras	0.02	Guiraud	0.01
Total de Stop Words	0.02	Dist. Euclidiana Bigrama	0.01
Total de sentenças	0.02	Cosseno Bigrama	0.01

Q3) A importância de contribuição dos atributos se repete em diferentes conjuntos de dados?

Alguns atributos se mantém para os diferentes conjuntos de dados. Os 10 principais atributos nos diferentes conjuntos de dados da língua inglesa foram: número de palavras diferentes, KNN com os melhores *scores*, total de caracteres, número de palavras longas, número de palavras, total de verbos, média do centroide global, total de conjunções, total de preposições e total de nominais. Agora, um modelo destes 10 melhores se aproxima-se em acurácia do modelo com 72 atributos? Rodando os mesmos experimentos apenas com os dez melhores atributos, a média da acurácia caiu 0.01 ponto, de $S\text{x}H$ 0.62 para $S\text{x}H$ 0.61.

5. Conclusões e trabalhos futuros

Neste trabalho, propôs-se um método de avaliação automática de ensaios na língua portuguesa baseado no refinamento de atributos de conteúdo (semânticos), de coerência e estatísticos de superfície para predizer a pontuação de um ensaio. O objetivo foi encontrar métodos robustos e com boa acurácia em relação a acurácia entre especialistas humanos. Utilizou-se uma arquitetura *pipeline* linear de 6 etapas: seleção de corpus, pré-processamento, extração de atributos, modelo de predição, validação, verificação da acurácia. Os experimentos produziram acurácias Kappa SxH 0.62 contra HxH 0.52 e acurácia Erro Médio SxH 0.92 contra HxH 0.89, resultados que superam a acurácia medida entre dois avaliadores humanos, indicando o potencial desta tecnologia para uso prático em ambientes virtuais de aprendizagem.

Os trabalhos futuros incluem expansão do corpus, busca da melhoria da acurácia e estudo de aprofundamento sobre a importância dos atributos e seus efeitos em diferentes conjuntos de dados.

Referências

aaaa.

Attali, Y. and Burstein, J. (2006). Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning and Assessment*, 4(3).

Burstein, J., Chodorow, M., and Leacock, C. (2004). Automated essay evaluation: The criterion online writing service. *Ai Magazine*, 25(3):27–27.

Burstein, J., Kukich, K., Wolff, S., Lu, C., and Chodorow, M. (1998). Computer analysis of essays. In *NCME Symposium on Automated Scoring*.

Carel, M. and Ducrot, O. (2001). O problema do paradoxo em uma semântica argumentativa. *Línguas e instrumentos lingüísticos*, 8:33–50.

Cheniti-Belcadhi, L., Braham, R., Henze, N., and Nejdl, W. (2004). A generic framework for assessment in adaptive educational hypermedia. In *ICWI*, pages 397–404.

Fleiss, J. L. and Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., and Landauer, T. K. (2013). Implementation and applications of the intelligent essay assessor. *Handbook of automated essay evaluation*, pages 68–88.

Grosz, B. J., Weinstein, S., and Joshi, A. K. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational linguistics*, 21(2):203–225.

Haley, D. T., Thomas, P., De Roeck, A., and Petre, M. (2007). Seeing the whole picture: evaluating automated assessment systems. *Innovation in Teaching and Learning in Information and Computer Sciences*, 6(4):203–224.

Laham, D., Foltz, P., and Landauer, T. (2000). The intelligent essay assessor. *IEEE Intelligent Systems*.

Mann, W. C. and Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Matthiessen, C. and Thompson, S. A. (1988). The structure of discourse and 'subordination'. *Clause combining in grammar and discourse*, 18:275–329.

Miller, D. I., Talbot, V., Gagnon, M., and Messier, C. (2013). Administration of neuropsychological tests using interactive voice response technology in the elderly: validation and limitations. *Frontiers in Neurology*, 4:107.

Page, E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243.

Palma, D. and Atkinson, J. (2018). Coherence-based automatic essay assessment. *IEEE Intelligent Systems*, 33(5):26–36.

Perkins, J. (2014). *Python 3 text processing with NLTK 3 cookbook*. Packt Publishing Ltd.

Rich, C. S., Schneider, M. C., and D'BROT, J. M. (2013). Applications of automated essay evaluation in west virginia. In *Handbook of Automated Essay Evaluation*, pages 121–145. Routledge.

Schultz, M. T. (2013). The intellimetric automated essay scoring engine-a review and an application to chinese essay scoring. *Handbook of automated essay scoring: Current applications and future directions*, pages 89–98.

Shermis, M. D. and Burstein, J. C. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Vajjala, S. (2018). Automated assessment of non-native learner essays: Investigating the role of linguistic features. *International Journal of Artificial Intelligence in Education*, 28(1):79–105.

Valenti, S., Neri, F., and Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education: Research*, 2:319–330.

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

Zupanc, K. and Bosnic, Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems*, 120:118–132.