Avaliação automática de redações na língua portuguesa baseada na coleta de atributos e aprendizagem de máquina

Silvério Sirotheau[1], Eloi Favero[2], João Santos[3], Simone Negrão[4], Marco Lima[5]

[1]Campus Universitário de Salinópolis - Universidade Federal do Pará (UFPA)
Rua Raimundo Santana Cruz S/N - São Miguel. CEP 68721-000 Salinópolis-PA-Brasil

[2]Programa de Pós-graduação em Ciência da Computação-PPGCC-UFPA

[3]Faculdade de Matemática - UFPA

[4]Campus Universitário de Castanhal - UFPA

[5]Faculdade de Computação - ICEN - UFPA

{silverio,favero,jcas,negrao}@ufpa.br, marconasc505@gmail.com

Abstract. *Virtual environments demand automatic evaluation methods for discursive questions. In the literature we find promising methods for texts in the English language, however, for Portuguese the studies are only preliminary. This research focuses on an approach of automatic evaluation of essays in Portuguese, based on the collection of features and the use of machine learning methods. In the experiments, 1000 essays from a public tender were used. In the collection of features, four dimensions were explored: Lexical, Syntactic, Content, and Coherence. As a result, we obtained the Kappa Square indexes (KQ) of 0.68 on the system against humans, versus a KQ of 0.56 on human against human.*

Resumo. *Ambientes virtuais demandam métodos de avaliação automática para questões discursivas. Na literatura encontramos métodos promissores para textos na língua inglesa, porém, para o português os estudos são apenas preliminares. Esta pesquisa foca numa abordagem de avaliação automática de redações em português, baseada na coleta de atributos e em métodos de aprendizagem de máquina. Nos experimentos utilizou-se 1000 redações de um concurso público. Na coleta de atributos explorou-se quatro dimensões: Léxica, Sintática, Conteúdo e Coerência. Como resultado foram obtidos índices Kappa quadrado (KQ) de 0.68 do sistema contra humanos versus um KQ de 0.56 de humano contra humano.*

1. Introdução

A avaliação é uma tarefa central no processo educativo, visto que é uma maneira de qualificar o conhecimento dos alunos em relação aos conceitos ensinados (Mohler e Mihalcea, 2009; Rodrigues e Araújo, 2012). O resultado da avaliação é também utilizado como uma ferramenta de *feedback* para orientação do aluno. Tradicionalmente a avaliação é feita pelo professor de forma manual, mas com o avanço tecnológico várias tarefas podem ser automatizadas (Bull e Mckenna, 2001; Amália *et al*., 2019). Sistemas

DOI: 10.5753/cbie.sbie.2020.1162

com essa finalidade são desenvolvidos pela pesquisa em avaliação automática de textos (AAT), relacionada à subárea Processamento de Linguagem Natural (PLN).

Em muitas plataformas de ensino, questões discursivas (ou ensaios) servem ao propósito de verificação e aferição da aprendizagem do aluno, em particular à capacidade de escrita e estruturação do discurso argumentativo (Page, 1966; Yang, 2012; Lee, 2014; Zupanc e Bosnic, 2015). Neste contexto, o desenvolvimento de abordagens que permitem automatizar a correção ganha relevância, por apresentar vantagens: possibilita *feedback* imediato; permite múltiplas submissões em tempo e lugar determinados pelo usuário; tem eficiência estável; e baixo custo.

Pesquisas em AAT iniciou-se na década de 60 (Page, 1966; Hearst, 2000; Noorbehbahani e Kardan, 2011), porém, somente na década de 90, com o uso de técnicas de PLN, houve um avanço considerável. Hoje, em termos de definir o escore de um ensaio, os sistemas já alcançam boa acurácia em relação aos avaliadores humanos (Shermis e Hammer, 2012). Porém, os sistemas deixam muito a desejar quanto ao *feedback* inteligente e orientador para o aluno, similar ao oferecido por um avaliador humano (Attali e Burstein, 2006; Haley *et al*., 2007).

Em provas discursivas avaliadas por dois especialistas humanos de forma independente, os avaliadores podem divergir nos seus escores. Por exemplo, considerando um único texto avaliado por dois especialistas humanos, numa escala de 0 a 10, se os seus escores divergem em 2 pontos, então dizemos que o percentual de concordância ou acurácia Humano *versus* Humano (H×H) é de 80%. Assim, num mesmo texto avaliado por um sistema e por um especialista humano, podemos medir a acurácia Sistema *versus* Humano (S×H). Um sistema possui bom desempenho quando a acurácia S×H é próxima ou superior à acurácia H×H. Para a literatura inglesa, vários estudos relatam boa aproximação e até superação de acurácia S×H contra H×H (Shermis e Hammer 2012; Srivastava *et al*., 2017).

Neste trabalho propomos uma abordagem para avaliar redações em português centrada na coleta de atributos. Serão considerados atributos relacionados a 4 dimensões: Léxica, Sintática, do Conteúdo e da Coerência. Recentemente surgiram também abordagens de coleta automática de atributos (mais de conteúdo) tais como *Convolutional Neural Networks* (Dong, Zhang e Yang 2017; Dasgupta *et al*., 2018). Estas abordagens têm alcançado e até superado as acurácias do escore final das abordagens de coleta de atributo manuais, porém elas têm restrições para dar *feedback* para o estudante. A metodologia que usa coleta de atributos permite enriquecer o escore com *feedback* orientador para o aluno, por exemplo, mostrando os erros gramaticais e/ou informando em qual grupo de atributos o texto teve deficiência (Zupanc e Bosnic, 2017). A meta é desenvolver uma acurácia SxH próxima da HxH, assim por exemplo, num evento real um dos avaliadores poderá ser substituído pelo computador.

Nesta investigação busca-se esclarecer algumas questões de pesquisas (QP) em torno do uso dos atributos nas dimensões: Léxica, Sintática, Conteúdo e Coerência. Qual a contribuição de cada uma das quatro dimensões (QP1-2-3-4) na acurácia final? Explora-se também o cruzamento entre as dimensões: (QP5) Qual a acurácia da combinação das dimensões duas a duas: léxica x sintática, léxica x conteúdo, léxica x coerência, sintática x conteúdo, sintática x coerência e conteúdo x coerência? Além disso, como a dimensão Conteúdo tem a maior influência explora-se conteúdo + léxica + coerência, conteúdo + sintática + coerência e conteúdo + léxica + sintática. (QP6) Com base nas quatros

dimensões, a acurácia final do método SxH alcança a acurácia dos avaliadores humanos HxH?

Além desta introdução este artigo está organizado da seguinte forma: A seção 2 apresenta os trabalhos relacionados. A seção 3 apresenta o corpus de estudo. A seção 4 apresenta a abordagem proposta. A seção 5 apresenta os resultados e discussão. Finalmente, a seção 6 apresenta a conclusão.

2. Trabalhos relacionados

Quanto às abordagens centradas na coleta de atributos, temos sistemas com algumas dezenas até com várias centenas: 60 atributos (*Intelligent Essay Assessor*); 90 atributos (Bokette); 150 atributos (*Semantic Automated Grader for Essays*); até mais de 400 atributos (IntelliMetric). Fonseca *et al*., 2018, utilizou 681 atributos num trabalho com textos em língua Portuguesa.

O *Intelligent Essay Assessor* (IEA) é um sistema de avaliação automática de ensaios. Ele usa Análise Semântica Latente (LSA) e PLN para coletar 60 atributos (Foltz, Laham e Landauer, 1999). Compreende um modelo de aprendizagem mecânica que induz a semelhança semântica de palavras baseada em grandes corpora de texto relevante para o domínio do tema e concentra-se mais em conteúdo do que em qualidade de escrita; possui também um módulo de *feedback.* Ele alcança uma acurácia KQ de 0.73 (Shermis e Hamner, 2012).

O Bookette (Rich *et al*., 2013) usa técnicas de PLN para extrair 90 atributos que descrevem a qualidade do texto. O sistema usa redes neurais para gerar escore e *feedback*. O sistema produz um *feedback* baseado no desempenho dos atributos, um *feedback* holístico e também comentários sobre as convenções de gramática e ortografia. O sistema alcança uma acurácia KQ de 0.70 (Shermis e Hamner, 2012).

Zupanc e Bosnic (2017) estenderam o sistema SAGE (*Semantic Automated Grader for Essays*) com atributos de coerência, baseados em janelas sequenciais (1/4 do texto, com passo de 10 palavras) de um ensaio, visualizado num espaço semântico. As medidas de coerência são o resultado das mudanças entre as janelas para estimar a coerência do texto. O sistema estendido com os novos atributos de coerências alcança uma acurácia KQ de 0.93.

O Intellimetric (Elliot, 2003) é um sistema que analisa elementos do discurso para formar um sentido semântico baseado em duas categorias: a) estrutura de atributos sintático - estruturais e mecânicos e b) conteúdo - discurso/retórica e atributos de conteúdo/conceito. O sistema usa múltiplas abordagens de aprendizagem de máquina (incluindo análise linear, analise bayesiana e LSA) para combinar os diferentes modelos em um único desfecho final (Rudner, Garcia e Welch, 2006). O sistema alcança uma acurácia KQ de 0.76 (Shermis e Hamner, 2012).

Para a língua portuguesa, Amorim e Veloso (2017) desenvolveram um experimento em AAT usando um conjunto de dados disponível na *web*, com 1840 redações (média de 300 palavras, do tipo ENEM) sobre 96 temas diferentes. Estas redações foram avaliadas em relação a 5 aspectos: Linguagem formal, foco no tema, organização do discurso, organização do argumento, solução proposta. A pesquisa se baseou em menos de 30 atributos, em classes como: sofisticação léxica, gramática e mecânica (erros léxicos e sintáticos e pontuação). Eles alcançaram uma acurácia KQ 0.42.

Também para a língua portuguesa, Fonseca *et al*., (2018) alcançaram uma acurácia KQ de 0.75 no escore geral em provas tipo ENEM, avaliadas nos mesmos 5 aspectos. Eles utilizaram uma abordagem de *Deep Neural Network* (DNN) e também baseada em um grande número de atributos (681) coletados manualmente. Os atributos foram coletados nas categorias: POS *tags*; POS *n-grams*; tokens *n-grams*; expressões específicas (agentes sociais, conectivos, etc.) e métricas de contagens. A abordagem baseada nos atributos foi vencedora na avaliação dos quatro primeiros aspectos. E a abordagem DNN foi vencedora para avaliar o quinto aspecto, a solução proposta.

Shermis e Hammer (2012) relataram que para dois humanos os escores variaram nas taxas de concordância de 0.61 a 0.85 medido por KQ; e nas pontuações das máquinas contra os humanos a concordância variou de 0.60 a 0.84 em KQ. Na língua portuguesa os estudos ainda são incipientes, portanto, este trabalho é mais uma contribuição para a maturidade desta área de pesquisa na nossa língua.

3. Corpus de estudo

O corpus da pesquisa foi composto por uma amostra de 1.000 redações de um concurso público do edital nº 26/2016 da Universidade Federal do Oeste do Pará. Estas redações passaram por um processo de digitalização manual onde não foi feito nenhum tipo de correção ortográfica e nem alterações nos aspectos gramaticais do texto original. Todas as redações foram previamente avaliadas por dois avaliadores humanos, recebendo uma pontuação inteira entre 0 e 10, com passo de 0.25; sendo que cada avaliador não conhece a pontuação do outro. Foram feitas verificações de discrepâncias: se as duas pontuações divergem por mais de um ponto, então um terceiro avaliador atribui uma pontuação para resolver a discrepância. A Figura 1 apresenta, num histograma, as classes de notas do corpus atribuídas pelos avaliadores humanos.

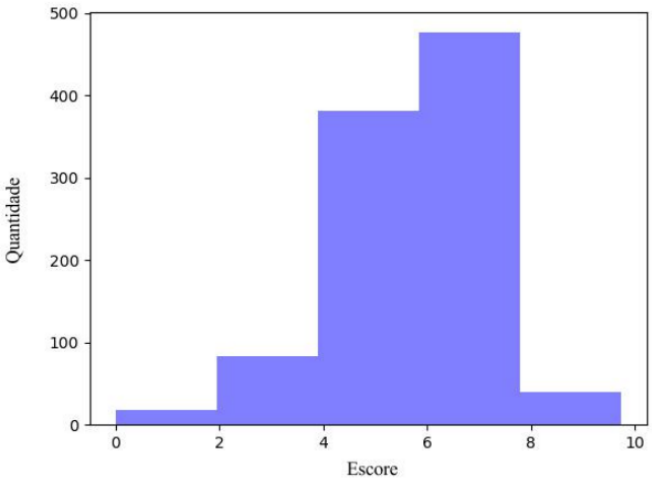

Figure 1. Histograma dos escores do corpus de redações, dos valores atribuídos pelos avaliadores humanos.

O escore final é média da avaliação de 3 competências: (i) tema está relacionada com compreender o tema e não fugir do que é proposto; (ii) coerência está relacionada com selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista e conhecimento dos mecanismos linguísticos necessários

para a construção da argumentação e; (iii) regras está relacionada com domínio da escrita formal da língua portuguesa.

4. Abordagem proposta

Na literatura existe uma tendência de uso de uma arquitetura de *pipeline* similar à da Figura 2 para sistema de AAT (Burrows, Gurevych e Stein, 2015). Ela contém 5 etapas: (1) preparação de corpus, (2) pré-processamento, (3) coleta de atributos, (4) modelo de predição e (5) avaliação. A saída de uma etapa é entrada para a próxima. A arquitetura *pipeline* é bem comum em vários campos da pesquisa e PLN, por exemplo, na extração informação, extração de relações e preenchimento de *templates* (Wachsmuth, Stein e Engels, 2011). Em cada uma dessas etapas os pesquisadores utilizam diferentes métodos e técnicas para no final gerar o escore de cada resposta.

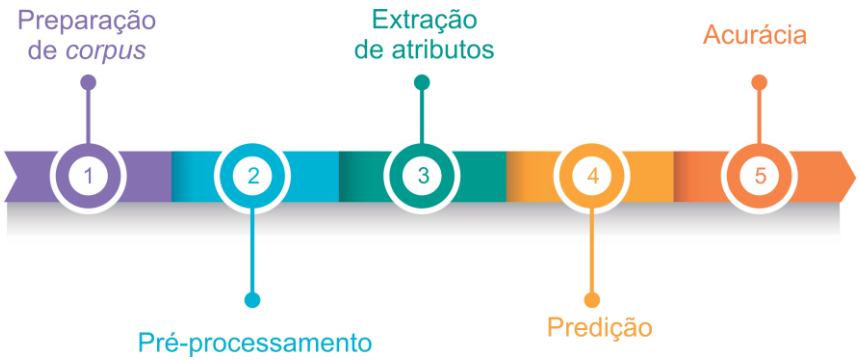

Figure 2. Arquitetura em *pipeline* para avaliação de textos composta de 5 etapas.

Na etapa de **preparação do corpus** as respostas das questões abertas são digitadas e organizadas numa coleção. Cada resposta da base tem pelo menos 2 escores de avaliadores humanos, permitindo a medida da acurácia HxH, para depois ser comparada com a acurácia SxH.

No **pré-processamento** busca-se uma representação "normalizada" do documento, deixando apenas a informação relevante para o processo de avaliação. As respostas foram separadas em sentenças e em seguidas vetorizadas em *tokens*. Após isso, três técnicas de pré-processamento foram aplicadas, com o uso biblioteca *Natural Language Toolkit* (NLTK): (1) Remoção de Caracteres Especiais, pontuação, acentuação e conversão para minúsculas (RCE); (2) Remoção de *stopword* (RSW) e; (3) Remoção de sufixos (*stemmer*) (RSU). As técnicas foram combinadas da seguinte forma: a) sem pré-processamento (-RCE, -RSW, -RSU); b) com remoção de caracteres especiais (+RCE, -RSW, -RSU); c) com remoção de caracteres especiais e *stopword* (+RCE, +RSW, -RSU) e; d) com remoção de caracteres especiais, *stopword* e aplicação de *stemmer* (+RCE, +RSW, +RSU). Em seguida, ainda nesta etapa, os *tokens* foram classificados morfologicamente com o etiquetador o Aelius (Alencar, 2010).

Na etapa de **extração de atributos,** procura-se abranger as principais classes de atributos citadas na literatura recente para a língua inglesa (Zupanc e Bosnic, 2017; Palma e Atkinson, 2018, Vajjala, 2018). Foram extraídos mais de 140 atributos agrupados em 4 dimensões (Tabela 1):

- **Léxica** descreve a coleta de atributos em um aspecto individual das palavras. Esta dimensão tem três principais categorias: (i) estatística de superfície, coleta estatística baseado em contagem de palavras; (ii) diversidade, coleta medidas que representam o quanto é diverso o vocabulário utilizado; (iii) legibilidade, mede o grau de facilidade da leitura do texto.
- **Sintática** descreve atributos que retratam o aspecto individual de cada sentença, compreende duas categorias: (1) número de cada *PoS tag* (*part-of-speech tagging*), como por exemplo, número de nomes (*noun*) e verbos (*verb*) (2) Erros, léxicos e sintáticos, conta o número de erros de sentenças mal formuladas, por exemplo, erros de concordância e pontuação.
- **Conteúdo** descreve atributos relacionadas com as medidas de similaridade entre as respostas dos alunos e a resposta de referência.
- **Coerência** descreve atributos relacionados à coerência do discurso de cada resposta (Zupanc e Bosnic, 2017).

Table 1. Exemplo de alguns atributos utilizados no experimento.

	Atributos
Léxica	*nº de caracteres, nº de diferentes palavras, nº de palavras, nº de palavras curtas, nº de palavras longas, nº médio de palavras, nº de stopword, nº de sentenças, comprimento de palavra mais frequente, nº de sílaba (...)*
Sintática	*nº de cada etiquetas (SR = ser, HV = haver, ET = estar, TR = ter, VB = verbo, Gênero, Número, Substantivo, Nome próprio, Pronomes, Preposição...), nº de diferentes etiquetas, nº de erros ortográficos (...)*
Conteúdo	*foram aplicadas as medidas (Cosseno e Distância Euclidiana) contra vetores de respostas, incluindo também as variações no tipo de pré-processamento (SSW, CST, CSW)*
Coerência	*foram utilizados quatro modelos combinando técnicas de pré-processamento e as duas medidas (Cosseno e Distância Euclidiana), geraram 3(a, b, c) x 2(min/máx. med.) x 2(cos, dist) x 3(pre).*

Na etapa de **predição**, utilizou-se o algoritmo de aprendizado de máquina supervisionado *Random Forest* que aceita na entrada de centenas de atributos em tarefas de regressão e/ou predição. Para este tipo de problema, onde temos mais de 100 atributos, o algoritmo tem duas vantagens: tem bom desempenho e retorna uma classificação da importância dos melhores atributos (Fernández-Delgado *et al.,* 2014). Para validação dos resultados utilizamos a abordagem *Cross-validation*, particionando o conjunto de dados em 5 *folds*; a acurácia coletada é a média dos 5 testes.

Na etapa de **acurácia**, seleciona-se as melhores combinações das etapas anteriores (por exemplo, diferentes combinações de pré-processamento) buscando maximizar o resultado final. Para medir a acurácia usou-se Kappa Quadrático - KQ (Fleiss e Cohen, 1973), que mede o grau de concordância entre duas classes com uma certa flexibilidade em relação à concordância exata. O KQ mede também a concordância parcial: se devia predizer 6, mas se resultou em 5, não é totalmente errado. Essa métrica geralmente varia de 0 (pouca concordância entre avaliadores) a 1 (concordância completa entre avaliadores). A escala de interpretação do KQ é: i) < 0.00 → "Pobre", ii) 0.00 - 0.20 → "Fraco", iii) 0.21 - 0.40 → "Razoável", iv) 0.41 - 0.60 → "Moderado", v) 0.61 - 0.80 → "Substancial" e vi) 0.81 - 1.00 → "Quase perfeito".

O KQ é calculado criando-se uma matriz de acordo com a equações 1 e 2. Cada célula da matriz, i.e., $O_{i,j}$ corresponde a uma resposta que pontuam i do avaliador humano e j do sistema. $W_{i,j}$ contém os pesos calculados conforme a Equação 1 e a matriz. $E_{i,j}$ contém as pontuações esperadas dos avaliadores humanos, obtidas pela multiplicação dos vetores de histograma das duas pontuações. Os subscritos em matriz $O_{i,j}$ correspondem ao número de ensaios que pontuam i do avaliador humano e j do sistema. No final do processo KQ e calculado de acordo com a equação 2.

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

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

5. Resultados e discussão

Realizou-se experimentos no corpus com 1.000 redações. Foram trabalhados com mais de 140 atributos, a grande maioria adaptados da língua inglesa. Em parte, pretendia-se verificar se os atributos para língua inglesa são adequados para o português. Aplicou-se a abordagem com os dados do corpus com o objetivo de maximizar o valor S×H buscando uma aproximação com H×H.

Em relação a questão de pesquisa QP1-2-3-4: Qual a contribuição de cada uma das quatro dimensões? Esta comparação foi feita considerando-se somente os atributos de uma dimensão para predizer o escore dos humanos. A Figura 3 mostra a contribuição das dimensões (léxica, sintática, conteúdo e coerência) presentando os valores de KQ (0.42, 0.46, 0.59 e 0.40), respectivamente, que são valores apenas moderados na interpretação do KQ. No entanto, o valor 0.59 para conteúdo se aproxima do 0.60 que já está na escala de interpretação "substancial". A melhor contribuição é do conteúdo.

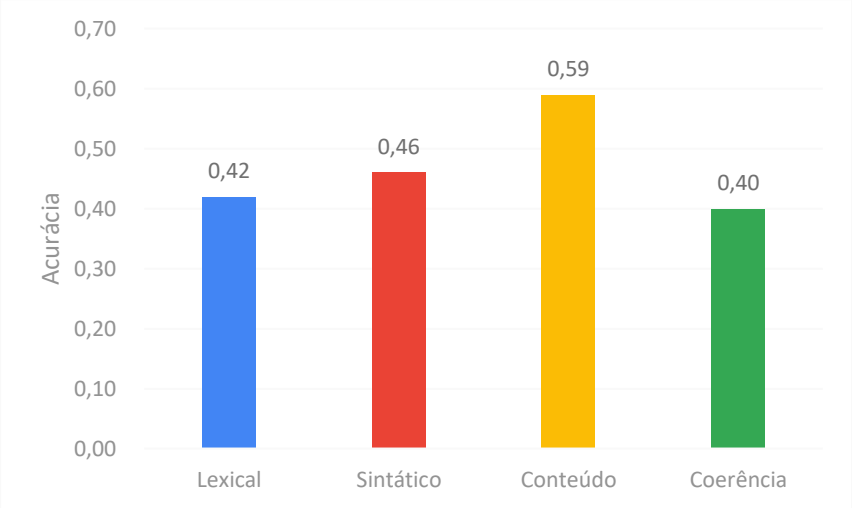

Figure 3. A contribuição de cada uma das dimensões (Léxica, Sintática, Conteúdo e Coerência) na acurácia final (Métrica: KQ).

Respondendo à questão de pesquisa (QP5), que busca explorar o cruzamento entre as dimensões. Qual a acurácia da combinação das dimensões duas a duas: Léxica x Sintática, ...? Com duas dimensões, a menor acurácia vem da combinação da dimensão léxica+sintática com KQ 0.56; já o melhor desempenho é com a combinação das dimensões de conteúdo+coerência com KQ 0.62 (Figura 4).

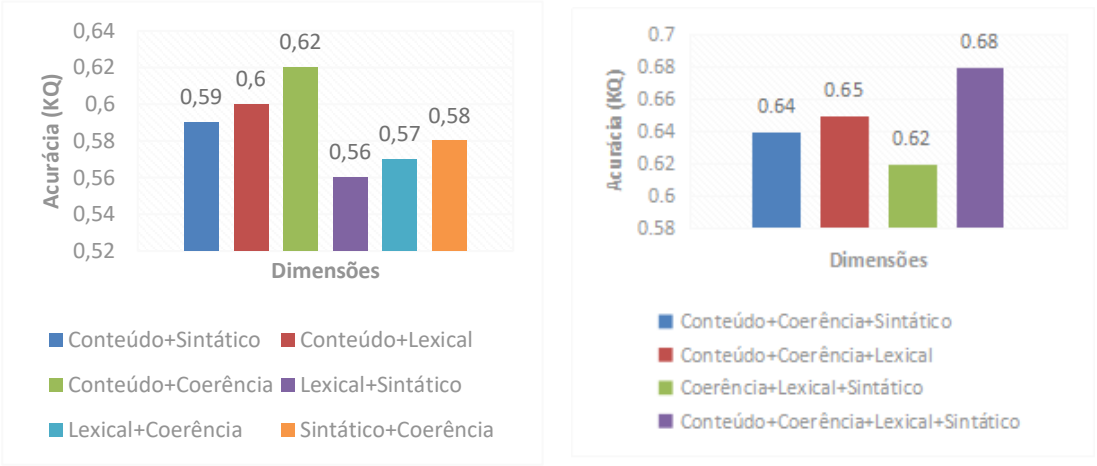

Figure 4. Explorando a combinação das dimensões 2 a 2 e 3 a 3 na contribuição da acurácia.

Além disso, como a dimensão Conteúdo tem a maior influência explora-se Conteúdo + Léxica + Coerência, Conteúdo + Sintática + Coerência e Conteúdo + Léxica + Sintática (Figura 4). Com três dimensões a pior combinação vem de conteúdo + léxica + sintática com 0.62 e a melhor vem de conteúdo + coerência + léxica, com KQ 0.65.

Combinando as quatro dimensões responde-se à questão de pesquisa (QP6): O método de avaliação alcança a acurácia dos avaliadores humanos? A combinação das quatro dimensões resultou uma acurácia final de KQ 0.68 contra a acurácia HxH com KQ 0.56. Portanto, o sistema supera a acurácia humana.

6. Conclusão

O objetivo deste trabalho foi desenvolver um método de avaliação automática de respostas discursivas para redação, coletando-se atributos em 4 dimensões. Foram classificadas numa espécie de taxonomia mais de 140 atributos. A maior parte delas veio de trabalhos relacionados da língua Inglesa as quais foram ajustadas para o Português.

Para realização dos experimentos utilizou-se uma arquitetura *pipeline* linear de 5 etapas. Utilizamos a técnica *Random Forest*, que permite a manipulação de um grande número de atributos além de retornar a relevância de cada atributo na etapa de classificação. Como resultado obtivemos um kappa quadrático (KQ) 0.68 SxH contra 0.56 HxH. Um resultado KQ de 0.68 é considerado "substancial". Este resultado superou a acurácia medida entre os avaliadores humanos. Com relação aos trabalhos prévios da língua portuguesa, esta acurácia supera o do trabalho de Amorim e Veloso (2017) com KQ 0.42 mas ainda é inferior ao valor de acurácia encontrado por Fonseca *et. al.,* 2018, com KQ 0.75. Este resultado mostra que esta tecnologia está alcançando um estado de maturidade para ser utilizada em aplicações práticas, por exemplo, em ambientes virtuais

de ensino. Como trabalho futuro pode-se estudar diversas frentes como, aprofundar a importância de cada atributo associado ao escore e/ou as competências e aplicar técnicas de *deep learning* na abordagem.

7. Referências

Alencar, L. F. de. (2010) Aelius: uma ferramenta para anotação automática de corpora usando NLTK. IX Encontro de Linguística de Corpus. Porto Alegre, PUCRS.

Amália, A. et al. (2019) Automated Bahasa Indonesia essay evaluation with latent semantic analysis. In: Journal of Physics: Conference Series. IOP Publishing.

Amorim, E. e Veloso, A. (2017). multi-aspect analysis of automatic essay scoring for Brazilian Portuguese. In Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics (pp. 94-102).

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

Bull, J. e Mckenna, C. (2001) A Blueprint for Computer Assisted Assessment. Taylor & Francis Editora.

Burrows, S, Gurevych, I. e Stein, B. (2015) The eras and trends of automatic short answer grading. International Journal of Artificial Intelligence in Education.

Dasgupta, I., Guo, D., Stuhlmüller, A., Gershman, S. J., e Goodman, N. D. (2018). Evaluating compositionality in sentence embeddings. arXiv preprint arXiv:1802.04302.

Dong, F., Zhang, Y. e Yang, J.(2017). Attention-based recurrent convolutional neural network for automatic essay scoring. In Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)(pp. 153-162)

Elliot, S. (2003). IntelliMetric: From here to validity. Automated essay scoring: A cross-disciplinary perspective, 71-86.

Fernández-Delgado, M., Cernadas, E., Barro, S. e Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. The Journal of Machine Learning Research, 15(1), 3133-3181.

Fleiss, J. L., e 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., Laham, D. e Landauer, T. K.(1999). The intelligent essay assessor: Applications to educational technology. Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1(2), 939-944.

Fonseca, E., Medeiros, I., Kamikawachi, D., e Bokan, A. (2018). Automatically grading brazilian student essays. In International Conference on Computational Processing of the Portuguese Language (pp. 170-179). Springer, Cham.

Haley, D. T. et al. (2007) Seeing the whole picture: evaluating automated assessment systems. ITALICS.

Hearst, M. A. (2000) The debate on automated essay grading. IEE Intelligeng Systems archive.

Lee, I. (2014). Teachers' reflection on implementation of innovative feedback approaches in EFL writing. English Teaching, 69(1), 23-40.

Mohler, M. e Mihalcea, R. (2009) Text-to-text semantic similarity for automatic short answer grading. EACL'09 - Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics.

Noorbehbahani, F. e Kardan, A. A. (2011) The automatic assessment of free text answers using a modified bleu algorithm. Computer & Education.

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

Palma, D. e Atkinson, J. (2018) Coherence-Based Automatic Essay Assessment. IEEE Intelligent Systems, v. 33, n. 5, p. 26-36.

Rich, C. S.; Schneider, M. C. e D'brot, J. M. (2013) Applications of automated essay evaluation in West Virginia. In: Handbook of Automated Essay Evaluation. Routledge. p. 121-145.

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

Rudner, L. M., Garcia, V. e Welch, C. (2006). An evaluation of IntelliMetric™ essay scoring system. The Journal of Technology, Learning and Assessment, 4(4)

Shermis, M. D., e Hamner, B. (2012). Contrasting state-of-the-art automated scoring of essays: Analysis. In Annual national council on measurement in education meeting (pp. 14-16).

Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. (2014) Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1):1929-1958.

Vajjalla, S. (2018) "Automated assessment of non-native learner essays: Investigating the role of linguistic features". International Journal of Artificial Intelligence in Education, v. 28, n. 1, p. 79-105.

Wachsmuth, H., Stein, B., e Engels, G. (2011). Constructing efficient information extraction pipelines. In Proceedings of the 20th ACM international conference on Information and knowledge management (pp. 2237-2240). ACM.

Yang, W. (2012). A study of students' perceptions and attitudes towards genre-based ESP writing instruction. The Asian ESP Journal, 8(3), 50-73.

Zupanc, K. e Bosnic, Z. (2015) Automated essay evaluation augmented with semantic coherence measures. IEEE International Conference on Data Mining (ICDM).

Zupanc, K. e Bosnic, Z. (2017). Automated essay evaluation with semantic analysis. Know.-Based Syst., 120(C):118 – 132. DOI: 10.1016/j.knosys.2017.01.006