Avaliação Automática de respostas discursivas curtas baseado em três dimensões linguísticas

Silvério Sirotheau[12], João Carlos dos Santos[1], Eloi Favero[12], Simone Negrão[1]

[1]Instituto Ciências Exatas e Naturais - Universidade Federal do Pará (UFPA)
Rua Augusto Corrêa, 01 - Guamá. CEP 66075 -110 - Belém - PA – Brasil

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

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

Abstract. *As the use of virtual environments grows, there is a need for a system of automatic evaluation of discursive answers. This paper proposes a method for automatic evaluation of discursive short answers based on a machine learning architecture. The predictive method is based on the collection of features (140) of similarity between texts in a taxonomy of three linguistic dimensions: lexical, syntactic and semantic. As a result, we obtained quadratic kappa 0.72 human x system (SxH) against 0.94 human x human (HxH) for the proof of biology and an accuracy 0.76 SxH against 0.58 HxH for the proof of geography.*

Resumo. *Com o crescimento do uso de ambientes virtuais cresce a necessidade de um sistema avaliador automático para respostas discursivas. Este trabalho propõem um método para avaliação automática de respostas discursivas curtas baseado numa arquitetura de aprendizagem de máquina de 5 etapas. O método preditivo é baseado na coleta de features (140) de similaridade entre textos numa taxonomia de três dimensões linguísticas: léxico, sintático e semântico. Como resultado obtivemos kappa quadrático 0.72 sistema x humano (SxH) contra 0.94 humano x humano (HxH) para a prova de Biologia e uma acurácia 0.76 SxH contra 0.58 HxH para a prova de Geografia.*

1. Introdução

Durante seu percurso escolar, o aluno passa por um processo de avaliação de ensino aprendizagem contínuo, cumulativo e sistemático. Mesmo diante das concepções pedagógicas mais modernas, as aplicações de avaliações compostas por questões discursivas têm forte relevância, pois avaliam resultados de aprendizagem do aluno, em particular sua capacidade de escrita e a compreensão de conceitos específicos em um determinado domínio (Zupanc e Bosnic, 2017; Page, 1966).

No entanto, a tarefa de correção manual desse tipo avaliação para um número grande de alunos é muito dispendiosa em termos de recursos humanos, tempo e dinheiro. Por exemplo, o Exame Nacional de Ensino Médio (ENEM) que é um processo seletivo para ingressar em instituições federais de ensino superior no Brasil, com mais de 6 milhões candidatos, possui em sua estrutura questões de texto dissertativo-argumentativo. Qual o tempo e o custo para avaliar mais de 6 milhões de textos?

Rababah e Al-Taani (2017) afirmam que a correção manual pode consumir muito tempo do professor e que sistemas computacionais podem auxiliar nesse tipo de tarefa.

DOI: 10.5753/cbie.sbie.2019.1551

Esse tipo de sistema contribui auxiliando o avaliador humano, liberando-o em parte da correção manual, assim ele pode direcionar sua atenção para focos mais específicos do processo de ensino-aprendizagem. Neste contexto, o desenvolvimento de algoritmos para automatizar a correção de respostas de questões discursivas torna-se muito relevante no processo ensino aprendizagem (Pérez et al. 2005).

No campo da avaliação automática de questões discursivas curtas existem duas principais linhas de pesquisa: (1) A primeira baseia-se em *corpus* e similaridade entre textos (Gomaa e Fahmy 2014; Pribadi et al., 2017) e; (2) A segunda baseia-se em métricas de similaridade entre redes de conceitos extraídos dos textos das respostas utilizando-se técnicas de aprendizagem de máquina e processamento de linguagem natural (PLN) (Mohler e Mihalcea 2009; Zupanc e Bosnic, 2017; Palma e Atkinson, 2018).

Na abordagem baseada em similaridade de texto o PLN é apenas superficial (coleta de *tokens*) enquanto que na abordagem de similaridade entre redes de conceitos são necessários métodos de PLN e de Aprendizagem de Máquina mais sofisticados (etiquetagem, resolução de pronomes, extração de entidades, entre outros).

Este trabalho propõem um método para avaliação automática de respostas discursivas curtas baseado numa arquitetura de aprendizagem de máquina de 5 etapas. O método preditivo é baseado na coleta de *features* (140) de similaridade entre textos numa taxonomia de três dimensões linguísticas: léxico, sintático e semântico. Uma de nossas contribuições é trabalhar com essas *features*, originalmente propostas para outras línguas, direcionando a pesquisa para a língua portuguesa. Busca-se alcançar um valor de acurácia próxima do obtido entre dois avaliadores humanos (*HxH*). Quando um sistema contrastado com humanos alcança valores de acurácia próximos a dos avaliadores humanos (*HxH*), ele se torna confiável para ser usado na correção das questões discursivas (Haley et al. 2007).

Esta pesquisa contribui para gerar tecnologia inovadora para avaliação automática de questões discursivas curtas. Esta tecnologia em ambientes virtuais de aprendizagem apresenta as vantagens: (i) *feedback* imediato para o aluno, mesmo numa turma muito grande de estudantes; (ii) baixo custo financeiro; permite múltiplas avaliações num desenvolvimento interativo da resposta; (iii) uniformidade na avaliação, pois independe do cansaço físico e emocional do avaliador; (iv) libera o professor da correção manual, permite que o mesmo direcione maior atenção para pontos específicos.

Este artigo está organizado da seguinte forma: Seção 2 apresenta trabalhos relacionados. Seção 3 apresenta a metodologia. Seção 4 apresenta resultados e discursão e a seção 5 apresenta a conclusão.

2. Trabalhos relacionados

As pesquisas sobre a avaliação automática de textos (respostas longas) iniciaram na década de 60 com o sistema PEG, com o foco em avaliar as habilidades do estilo de escrita dos estudantes (Page 1966). Posteriormente outras iniciativas surgiram a partir dos anos 90, com o surgimento de técnicas de PLN proporcionando um avanço considerável neste campo como *E-rater* (Burstein et al. 1998) e *Intellimetric* (Learning 2000).

Esforços mais recentes vêm alcançando uma acurácia bem próxima a medida entre avaliadores humanos. Leacock e Chodorow (2003) descrevem um mecanismo de pontuação de respostas discursivas curtas a partir de resposta de referência de

especialistas, combinando *features* sintáticas de uma resposta do aluno (sujeito, objeto e verbo) com um conjunto de respostas de referência. Trabalhou com um *corpus* de 16.625 respostas, alcançando uma acurácia de concordância de 84% contra os avaliadores humanos.

Mohler e Mihalcea (2009) exploraram técnicas não supervisionadas de aprendizagem de máquina para a avaliação automática de respostas curtas. Foram combinadas medidas baseadas em conhecimento do *WordNet* e *Latent Semantic Analysis* (LSA). Eles alcançaram uma correlação de 0.50 ($S \times H$) contra uma de 0.64 de ($H \times H$).

Gomaa e Fahmy (2014) utilizaram diversas métricas de similaridades (*String-based similarity*, *Corpus-based similarity*, *Knowledge-based similarity*, *Hybrid similarity measures* e *Sentence-level semantic similarity*) como entrada do método de classificação; numa base de 610 respostas curtas de estudantes avaliadas numa escala de 0 a 5, obtiveram uma correlação de 0.68 ($S \times H$) contra a correlação de 0.86 ($H \times H$).

Rodrigues e Araújo (2012) exploraram técnicas de PLN com uma etapa de tradução de frases para formas canônicas (listas de palavra *vs.* 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.

Galhardi et al., (2018) apresentam uma nova base de dados para respostas curtas em português, com uma abordagem composta por 4 grupos de *features* (*bag of n-grams*, similaridade lexical, similaridade semântica e estatísticos) obtendo os resultados usando *Extreme Gradient Boosting Classifer* e *cross-validation,* alcançando uma acurácia de 69% e uma concordância *Kappa* de 0.54.

2.1 Questões de pesquisa

No levantamento bibliográfico sobre respostas discursivas do tipo curta, foram levantadas algumas questões (Q): **I -** Dentre as várias técnicas de pré-processamento (Burrows et al. 2015), neste trabalho são utilizadas três: de superfície (ex. remoção de pontuação), léxico (ex. correção ortográfica e remoção de *stop word*), morfológico (ex. *stemmer*). (Q1) o pré-processamento influência na acurácia final nesse tipo de abordagem? **II -** Vajalla (2018) afirma que pouco se conhece sobre quais *features* linguísticos são bons preditores. (Q2) quais os melhores *features* preditores para a língua portuguesa em questões de respostas discursivas curtas? (Q3) A importância de contribuição dos *features* se repete em diferentes conjuntos de dados?

3. Metodologia

A abordagem é centrada em uma arquitetura *pipeline* que contém 5 etapas: (1) seleção de *corpus*, (2) pré-processamento, (3) extração de *features*, (4) modelo de predição e (5) acurácia (ver Figura 1).

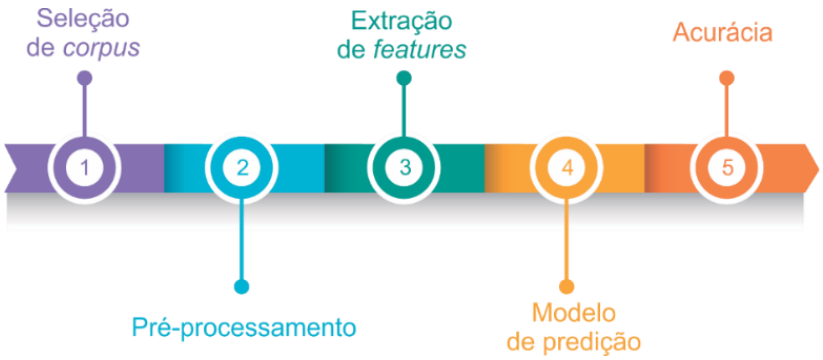

FIGURA 1. Arquitetura *pipeline* para avaliação de textos curtos.

Na etapa de seleção de *corpus*, foram selecionados dois conjuntos de dados de respostas curtas para língua portuguesa, relacionados a duas questões oriundas de uma prova de vestibular. Uma questão é de Biologia com 130 respostas e a outra é de Geografia com 229 respostas.

Na etapa de pré-processamento, as respostas foram vetorizadas em sentenças e em seguidas separadas em *tokens*. Após isso, três técnicas de pré-processamento foram utilizadas: (1) Remoção de Caracteres Especiais, pontuação, acentuação e conversão de letras maiúsculas em letras minúsculas (RCE); (2) Remoção de *stop word* (RSW) e; (3) Remoção de sufixos (*stemmer*) (RSU). Para o pré-processamento utilizamos a biblioteca *Natural Language Toolkit* (NLTK), onde 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 *stop word* (+RCE, +RSW, -RSU) e; d) com remoção de caracteres especiais, *stop word* e aplicação de *stemmer* (+RCE, +RSW, +RSU). Em seguida, os *tokens* foram etiquetados morfologicamente para classificação conforme suas categorias gramaticais, para isto utilizamos o Aelius (Alencar, 2010).

Na etapa de extração de *features* (atributos, características ou variáveis do texto), procuramos abranger todos os principais atributos na literatura recente (Zupanc e Bosnic, 2017; Palma e Atkinson, 2018; Vajjala, 2018) (ver tabela 1). Foram extraídas 140 *features* agrupadas em 3 dimensões: léxica, sintática e semântica.

Dimensão Léxica. Coleta *features* que descrevem o aspecto individual das palavras. Nesta dimensão temos 4 principais categorias: (1) estatística de superfície, coleta estatísticas baseado em contagem de palavras. (2) diversidade, coleta medidas que representam o quanto é diverso o vocabulário utilizado. (3) redabilidade, mede o grau de facilidade da leitura do texto. (4) Erro, número de erros ortográficos.

Dimensão Sintática. Coleta *features* 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) Erro Léxico e Sintático, conta o número de erros de sentenças mal formuladas, por exemplo, erros de concordância e pontuação.

Dimensão Semântica. Coleta *features* que descrevem os aspectos que estão relacionados ao conteúdo do texto, por exemplo, medidas de similaridade entre a resposta do aluno e a resposta de referência. E, coleta também *features* que descrevem os aspectos relacionados à coerência textual, tanto local dentro de uma resposta como global em relação as várias respostas.

TABELA 1. Taxonomia das *features* para avaliação automática de textos na língua portuguesa, para respostas curtas (parte 1/2).

Léxica	Estatística de Superfície	n° de caracteres, n° de diferentes palavras, n° de palavras, n° de palavras curtas, n° de palavras longas, n° média de palavras, n° de *stop word*, n° de sentença, n° comprimento de palavra mais frequente.
	Diversidade	*Type-token-ratio* – TTR, Guiraud's *index*, *Yule*'s K, *the D estimate*, hapax legomena.
	Redabilidade	*Gunning Fox Index*, *Flesch Kincaid grade level*, *Dale-Chall readability formula, autometed readability index*, LIX, *word variation index*, *nominal ratio*, SMOG-*index*
	Erro	n° de erros de ortográficos
Sintático	Número de cada PoS *tags*	Número de diferentes PoS *tags*, Número de *tags* por categoria sintática: SR=ser, HV=haver, ET=estar, TR=ter, VB=verb (-I=*imperative*, -P=presente, -SP=presente subjuntivo, -D=past, -RA=fonema inflexional, -SD=*past subjunctive*, -R=*future*, -SR=*future subjunctive*, -G=*gerund*, -PP=*perfect participle*, -NA=*agrement particle*), *Agreement Particle* (*genre*(*none*=masc, -F=fem, -G=*double gender*), *number*(none=sing, -P=plural)) , N (*noun*) NPR (*proper noun*) PRO (*pronouns*) P+PRO (Prep+*Pronoun*) PRO$ (*possessive*) CL (*clitics*) D (*determine*) DEM (*demonstrative*) ADJ (*adjective*) ADV (*adverbs*) Q (*quantifier*) CONJ (*conjunction*) C (*subordinating conjunction*) WPRO (*relative*) WQUE (*interrogative*) WD (*interrogative determiners*) P (*preposition*)
	Erro	n° de erros de pontuação

Na coleta de *features* de conteúdo uma resposta de aluno é contrastada com a resposta de referência, comum no uso de *n*-gramas (uni e bi) com as medidas de distância euclidiana e cosseno. Utilizamos também métodos de ponderação local e global dos textos como *tf-idf*. Tipicamente a resposta de referência é formada a partir de um conjunto das respostas mais bem avaliadas.

Por outro lado, alguns autores sugerem que se pode utilizar também respostas de referência baseadas em agrupamentos feitos em relação ao escore (Zupanc e Bosnic, 2017). Baseado nisso, considerando os escores na faixa de 0 a 6, criamos 7 vetores resposta de referência um para cada valor de escore. Aqui aplicamos as medidas (distância euclidiana e cosseno) contra estes vetores de respostas, incluindo também as variações no tipo de pré-processamento; disso resultou 66 *features* de conteúdo, que muitas delas estão nas de maior relevância. Ainda na dimensão semântica avaliamos à coerência do texto, que descreve o fluxo de informação dentro texto. Para isso, utilizamos uma abordagem baseada em janelas sobrepostas (Zupanc e Bosnic, 2017; Palma e Atkinson, 2018). Utilizamos 4 modelos que geraram 66 *features*, conforme tabela 2.

TABELA 2. Taxonomia das *features* (parte 2 / 2).

Semântico**	Conteúdo	Similaridade* Cosseno e distância euclidiana com resposta de referência	Similaridade com Texto Fonte (Pré: SSW, CST, CSW) (Med: Cosseno e Distância Euclidiana)
		Similaridade e distancia contra as faixas dos escores	Similaridade (nível: 0, 1, 2, 3, 4, 5, 6) (Pré: SSW, CST, CSW) (Med:Cosseno e Distância Euclidiana)
		Soma ponderada de todos os valores de	Correlação Ponderada (Pré: SSW,

		correlação baseada nos valores de cosseno e distância euclidiana	CST, CSW) (Med: Cosseno e Distância Euclidiana)
	Coerência	Distancias entre duas janelas contiguas	Min, med, max
		Distancias de todas janelas contra todas	
		Centro local, todas janelas contra o centro local	
		Centro global, todas janelas contra o centro global	

Similaridade* Cosseno e distância euclidiana: na verdade o cosseno é uma medida de similaridade, enquanto que a distância euclidiana é uma medida de dissimilaridade. Para tornar as duas medidas como similaridade consideramos 1/distancia euclidiana.

** a quantidade de *features* é dado pela multiplicação simples de cada grupo de itens, por exemplo, para conteúdo temos 7x3x2=42.

Na etapa de predição, utilizamos o algoritmo *Random Forest*, que como um método de aprendizado de máquina supervisionado permite a combinação de centenas de *features* em tarefas de regressão e/ou predição. Ele cria um conjunto de árvores de decisão, onde cada árvore é treinada por um subconjunto diferente de dados de treinamento. Para este tipo de problema, onde temos um grande número de *features*, mais de 100, o algoritmo *Random Forest* tem bom desempenho (Fernández-Delgado et al. 2014). Para validação 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, procura-se selecionar as melhores combinações das etapas anteriores buscando maximizar a precisão. Para medir a acurácia utilizamos *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). Caso a concordância entre os avaliadores seja abaixo do mínimo esperado, essa métrica também pode resultar em valores negativos.

O KQ é calculado criando-se uma matriz de acordo com a equações 1 e 2. Neste caso, a matriz O contém as pontuações, de tal modo que a classificação i é dada pelo avaliador humano e j dada pelo modelo. $W_{i,j}$ contém os pesos como derivado na Equação 1 e a matriz E 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 respostas que pontuam i do avaliador humano e j do sistema.

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

No final do processo KQ e calculado como:

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

A interpretação dos resultados de KQ, entre 0 e 1, entre pouca e muita concordância, pode ser um tanto subjetiva. Portanto citamos uma interpretação

recomendada por Landis e Koch (Landis e Koch, 1977) que considera seis faixas de valores: 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".

4. Resultados e discussão

O nosso corpus de pesquisa foi constituído por uma coleção de respostas a duas questões discursivas que constam no edital 016/2007 do vestibular da Universidade Federal do Pará. De um universo de mil folhas de respostas foram selecionadas as duas questões com mais folhas de respostas preenchidas: Biologia com 130 respostas e Geografia com 229 respostas. O candidato escolhia um subconjunto das 26 questões que iria responder, por isso não temos mil respostas por cada questão. A questão de Biologia possui em médias 28 palavras por resposta e a de Geografia 74 palavras. Cada resposta possui a nota de dois avaliadores humanos, portanto podemos calcular a acurácia de acerto entre eles ($H \times H$).

Aplicou-se a abordagem nas bases com a meta de maximizar o valor $S \times H$ buscando uma aproximação com $H \times H$. A tabela 3 apresenta os resultados em KQ.

TABELA 3. Resultados das respostas curtas de Biologia e de Geografia.

Base de dados	*SxH*	*HxH*
Biologia	0.72	0.94
Geografia	0.76	0.58

Esta tabela consolida os resultados que são promissores. Para Biologia o *HxH* foi de 0.94, na interpretação de KQ acima é uma concordância quase perfeita. O sistema alcançou um valor *SxH* 0.72, que é uma concordância substancial. Na questão de Geografia o *HxH* foi de 0.58 que é uma concordância moderada, no entanto o sistema alcançou um *SxH* 0.76, que é uma concordância substancial, superando o *HxH*.

Em relação à questão de pesquisa levantada (Q1), onde Burrows et al., (2015) relatam várias técnicas de pré-processamento utilizadas em processamento de texto. 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 (*stemmer*) (+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 tabela 4 temos os resultados obtidos para Biologia e Geografia considerando as variações nas técnicas de pré-processamento.

TABELA 4. Pré-processamento das respostas curtas de Biologia e de Geografia.

	Biologia		**Geografia**	
humano vs. humano	0.94		0.58	
Média de palavras por resposta	28.48		74.56	
Sistema vs. Humano	Cont	Lex+Sint+Cont	Cont	Lex+Sint +Cont
-RCE, -RSW, -RSU	0.65	0.64	0.70	0.70
+RCE, -RSW, -RSU	0.70	0.70	0.74	0.71
+RCE, +RSW, -RSU	0.64	0.64	0.66	**0.76**
+RCE, +RSW, +RSU	**0.71**	**0.72**	**0.73**	0.69

Conforme a Tabela 4, as diferentes técnicas de pré-processamento apresentam diferentes valores de acurácia. No entanto, as diferenças são bem significativas dentro de cada base, sendo a diferença do menor para o maior valor 0.08 em Biologia e 0.10 para Geografia, o que responde à questão de pesquisa Q1. Considerando estes valores é importante ter a etapa de pré-processamento nas abordagens de avaliação automática de respostas curtas.

Na discussão da questão de pesquisa Q2, sobre quais são as melhores *features* preditores para a língua portuguesa em questões de respostas discursivas curtas? Na tabela 5 apresentamos as principais *features* por ordem de importância. Partimos com um conjunto de mais de 140 *features*, e utilizamos o método *random forest* para predição e seleção da importância das *features*.

TABELA 5. Importância dos *features* nas respostas curtas de português.

Features	**Importância**	***Features***	**Importância**
Cosseno Escore 4	0.23	Distância Euclidiana Escore 0	0.06
Cosseno Escore 6 sem *Stop Word (SW)*	0.13	Cosseno Escore 6	0.05
Cosseno Escore 5	0.12	Cosseno Escore 3	0.05
Número de caracteres	0.09	Cosseno Escore 3 com *SW*	0.05
Cosseno Escore 5 sem *SW*	0.09	Número de palavras	0.05
Cosseno com texto fonte e com *SW*	0.07	Cosseno Escore 4 com SW	0.04
Número de *stop word*	0.06	Número de pronomes	0.04
Número de palavras longas	0.06	Número diferente de palavras	0.03

Em relação à questão de pesquisa (Q3) A importância de contribuição dos *features* repete-se nos diferentes conjuntos de dados? Na tabela 6 selecionamos os melhores *features* de cada base e verificamos que as medidas de cosseno e distancia euclidiana por faixa de escore são as principais *features* das duas bases. Por outro lado, as outras *features* mais relevantes são os léxicos de estatísticas de superfície, tais como número de palavras e número de palavras diferentes e número de palavras longas.

TABELA 6. Resultado da importância dos *features* em cada base de dados.

	Biologia		**Geografia**	
N°	***Features***	**Importância**	***Features***	**Importância**
1	cosseno escore 6 sem *stop word*	0.13	cosseno escore 4	0.23
2	cosseno escore 5	0.11	euclidiana escore 0	0.06
3	cosseno escore 5 sem *stop word*	0.09	cosseno escore 3	0.05
4	número de caracteres	0.09	número de *stop word*	0.05
5	cosseno com texto fonte	0.07	cosseno escore 3 com *stop word*	0.04
6	cosseno escore 6	0.05	cosseno escore 4 com *stop word*	0.04
7	número de palavras longas	0.04	número de palavras	0.03
8	cosseno texto fonte sem *stop word*	0.03	número de palavras diferentes	0.03

9	número de pronomes	0.03	cosseno escore 2	0.02
10	cosseno escore 4 sem *stop word*	0.02	cosseno escore 2 com *stop word*	0.02

5. Conclusão

O objetivo deste trabalho é desenvolver um método de avaliação automática de respostas discursivas curtas baseadas na similaridade entre textos, coletando-se *features* em três principais dimensões: léxico, sintático e semântico. Foram classificadas numa espécie de taxonomia mais de 140 *features*. 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: seleção de corpus, pré-processamento, extração de *features*, modelo de predição e acurácia.

A partir dos valores das *features* coletadas, o objetivo é predizer o valor do escore de cada resposta com uma acurácia próxima aquela medida entre dois avaliadores humanos. Utilizamos a técnica *Random Forest*, que permite a manipulação de um grande número de *features* além de retornar a relevância de cada *features* na etapa de classificação. Como resultado obtivemos um kappa quadrático (KQ) 0.72 *SxH* contra 0.94 *HxH* para a prova de Biologia e um valor *SxH* 0.76 contra *HxH* 0.58 para a prova de Geografia. Um resultado KQ de 0.72 é considerado "substancial" mesmo sendo inferior ao coletado entre dois avaliadores humanos. Por outro lado, na prova de Geografia o sistema com 0.76, também "substancial", supera a acurácia medida entre os dois avaliadores humanos que foi de 0.58, valor "moderado". Este resultado mostra que esta tecnologia está alcançando um estado de maturidade para ser utilizada em ambientes virtuais de ensino.

7. Referencias

Alencar, L. F. (2010) "Aelius: uma ferramenta para anotação automática de corpora usando o NLTK", Anais do IX Encontro de Linguística de Corpus, PUCRS, Porto Alegre, v. 8.

Burrows, S., Gurevych, I. and 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. et al. (1998) "Automated scoring using a hybrid feature identification technique", In Proceedings of the 17th international conference on Computational linguistics-Volume 1, pages 206–210. Association for Computational Linguistics.

Fernández-Delgado, M. et al. (2014) "Do we need hundreds of classifiers to solve real world classification problems?". The Journal of Machine Learning Research, v. 15, n. 1, p. 3133-3181.

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, v. 33, n. 3, p. 613-619.

Galhardi, L. et al. (2018) "Portuguese Automatic Short Answer Grading". In: Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE). p. 1373.

Gomaa, W. H. and Fahmy, A. A. (2014). "Automatic scoring for answers to arabic test questions". Computer Speech & Language, 28(4):833–857.

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.

Landis, J. R. and Koch, G. (1977) "The measurement of observer agreement for categorical data". biometrics, p. 159-174.

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

Mohler, M. and 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.

Page, E. B. (1966). "The imminence of... grading essays by computer". The Phi Delta Kappan, v. 47, n. 5, p. 238-243.

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

Pérez, D., Alfonseca, E., Rodríguez, P., Gliozzo, A., Strapparava, C., and Magnini, B. (2005). "About the effects of combining latent semantic analysis with natural language processing techniques for free-text assessment". Revista signos, 38(59):325–343.

Pribadi, F. S., Adji, T. B., Permanasari, A. E., Mulwinda, A., and Utomo, A. B. (2017). "Automatic short answer scoring using words overlapping methods". In AIP Conference Proceedings, volume 1818, page 020042. AIP Publishing.

Rababah, H. e Al-Taani, A. T. (2017) "An automated scoring approach for Arabic short answers essay questions". In: 8th International Conference on Information Technology (ICIT). IEEE, p. 697-702.

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

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.

Zupanc, K. and Bosnic, Z. (2017) "Automated essay evaluation with semantic analysis". Knowledge-Based Systems, v. 120, p. 118-132.