SCATeDi: Sistema Inteligente para Avaliação de Desempenho Escolar em Avaliações Discursivas

Ricardo F. Vilela[1], Pedro Henrique D. Valle[1], Raphael J. Muniz[1], Wênder A. Lima[1], Ana Carolina G. Inocêncio[1], Paulo Afonso P. Junior[1]

[1]Curso de Ciência da Computação – Universidade Federal de Goias (UFG)
Caixa Postal 01 – 75.801-615 – Jataí – GO – Brasil

ricardo.ufg@live.com,
pedrohenriquevalle@hotmail.com,rafhael_mj12@hotmail.com,
walipij@gmail.com,anacarolina.inocencio@gmail.com,
pauloafpjunior@gmail.com

Abstract. *The manual correction of essay questions is a non-trivial task. The time availability of most educators is limited thus hampering the teaching and learning of students. The time spent to fix up an evaluation is large, and often the educator can not maintain the same level of correction for all events. This paper aims to present an expert system based on fuzzy logic for automatic correction of essay questions, called SCATeDi. An evaluation of the system was performed with fifteen questions arising from three students of geography. The results were satisfactory, since the SCATeDi reached an error rate just between 0.076 and 0.506 points.*

Resumo. *A correção manual de questões discursivas não é uma tarefa trivial. A disponibilidade de tempo da maioria dos educadores é limitada prejudicando assim o ensino e a aprendizagem dos alunos. O tempo gasto para se corrigir uma avaliação é alto, e muitas das vezes o educador não consegue manter o mesmo nível de rigidez para a correção para todas as provas. O presente artigo tem como objetivo apresentar um sistema especialista baseado em Lógica Fuzzy para correção automática de questões discursivas, denominado SCATeDi. Uma avaliação do sistema foi realizada com quinze questões advindas de três alunos do curso de geografia da EJA (Educação de Jovens e Adultos). Os resultados apresentaram-se satisfatórios, uma vez que o SCATeDi conseguiu atribuir notas às avaliações destes alunos com erros que variam entre 0,076 e 0,506 pontos por avaliação.*

1. Introdução

O processo de avaliação de discentes tem o intuito de descrever as aptidões dos alunos que passam por uma série de testes que devem ser avaliados de forma justa, sem o favorecimento de indivíduos. Os testes podem ser classificados em três tipos [Bezerra 2008]: **i) objetivos**: são aqueles que utilizam questões de múltipla escolha para avaliar o discente, porém não captam informações de como ele chegou até a resposta; **ii) orais**: submetem os estudantes a perguntas que devem ser respondidas em pronta exatidão, que diferentemente dos testes objetivos, podem avaliar se os estudantes possuem conhecimento prévio sobre o assunto, porém, é uma forma constrangedora para alguns deles, o que pode implicar em uma realização inadequada do processo de avaliação dos mesmos; e **iii) discursivos**: são baseados em questões dissertativas, as quais o aluno deve responder em forma de texto. Segundo Novak (1984), testes com questões

ISSN: 2316-6541

dissertativas demonstram de maneira mais clara o raciocínio do aluno, entretanto tal tipo de teste exige um maior esforço por parte dos docentes.

Perrenoud (1999) cita como problema que pode ocorrer durante a correção de testes dissertativos a falta de coerência do professor ao corrigi-los, ou seja, o avaliador pode entender de diversas formas o resultado proposto pelo discente, ocasionando em uma má avaliação do conhecimento do mesmo. Outro problema é a dificuldade em manter um mesmo nível de rigidez durante a correção de todas as avaliações.

Segundo Caldas e Favero (2009), a avaliação automática de questões fornece algumas vantagens em relação à avaliação manual, tais como: i) impõe menor carga de esforço ao docente, permitindo que este dedique-se a uma avaliação mais qualitativa e individual de cada estudante; ii) oferece maior visibilidade ao docente do desempenho dos alunos através de relatórios avaliativos, permitindo ao professor identificar problemas, como a deficiência de um aluno em relação a um tópico ministrado de forma mais rápida e tomar decisões para solucioná-las de forma mais eficiente; e iii) oferece rápido retorno dos resultados alcançados pelos estudantes em uma avaliação dissertativa, característica importante principalmente em ambientes EAD (Educação a Distância), no qual o professor não está sempre disponível. Além disso, a avaliação automática pode ajudar a minimizar os problemas da incidência de incoerências ou imparcialidade na correção das questões de uma avaliação.

Diversos trabalhos vêm sendo realizados para avaliação automática de desempenho escolar, porém muitos deles ou são baseados em testes objetivos (com questões de múltipla escolha) [Moodle 2012], [Nozawa e Oliveira 2006], [Abrão e Rayel 2005], [Perrenoud 1999], ou são baseados em testes discursivos, porém exigem padrões de respostas que limitam a expressividade dos discentes ao responder estes testes. Por exemplo, Caldas e Favero (2009) propõe a avaliação automática de mapas conceituais, porém a utilização deste tipo de modelo de resposta força os estudantes a seguirem uma estrutura pré-determinada, que por muitas vezes não capta o total conhecimento do aluno.

Entende-se que o processo de correção de questões discursivas possui certos padrões que são utilizados por especialistas, no caso os docentes, e que podem ser mapeados para um sistema especialista, possibilitando assim a correção automática destas questões. Sistemas especialistas são sistemas computacionais que simulam o raciocínio desenvolvido por especialistas em determinados contextos. Por exemplo, um professor, como especialista em sua área, saberá identificar se resposta do aluno contém as informações importantes que a classificam como correta. Para chegar a esta identificação, o avaliador utilizará de certo conhecimento especializado que o permitirá tomar decisões e atribuir uma nota à questão do aluno, como: i) a resposta do aluno possui as informações necessárias para ser considerada correta?; ii) a resposta do aluno possui as informações necessárias, porém também possui muita informação desnecessária e desvinculada do objetivo da questão?; e iii) o desempenho geral da turma foi bom, muito bom, ruim ou péssimo? Desse modo, estratégias de Inteligência Artificial, como a Lógica *Fuzzy*, podem ser adequadas à modelagem deste sistema, uma vez que as variáveis que o professor pode utilizar para atribuir uma nota a uma questão podem não ter uma fronteira bem definida.

O objetivo do presente trabalho é apresentar um sistema especialista baseado em Lógica *Fuzzy*, denominado SCATeDi (Sistema para Correção Automática de Testes

ISSN: 2316-6541

Discursivos) para correção automática de testes discursivos. Para realizar a correção, o sistema recebe as respostas dos estudantes, juntamente com o gabarito do professor e calcula um conjunto de métricas relacionadas à quantidade de termos corretos utilizados pelos estudantes e a precisão de suas respostas para atribuir uma nota que esteja o mais próximo da nota almejada pelo professor.

Este artigo está organizado da seguinte forma: na Seção 2 são apresentados sucintamente os conceitos básicos da Lógica *Fuzzy*. Na Seção 3 alguns dos trabalhos relacionados são discutidos. Na Seção 4 é apresentada a arquitetura e a modelagem da ferramenta SCATeDi. Na Seção 5 é apresentado um estudo experimental realizado com o objetivo de verificar a aplicabilidade e a eficiência da ferramenta proposta neste artigo. Por último, na Seção 6 estão as considerações finais e propostas de trabalhos futuros.

2. Lógica *Fuzzy*

A lógica *fuzzy* foi criada em 1965 por Lotfi A. Zadeh [Zadeh 1965], professor do Departamento de Engenharia Elétrica da Universidade da Califórnia em Berkeley e pode ser entendida como uma generalização da lógica clássica que admite infinitos valores lógicos intermediários entre a falsidade e a verdade.

Segundo Souza (2004), a diferença entre a lógica *fuzzy* e a lógica clássica é que na primeira o resultado de uma proposição pode assumir mais de dois valores distintos, diferentemente da lógica clássica, que pode assumir apenas dois valores (verdadeiro ou falso). Esses resultados não são expressos de forma bem definida, mas linguisticamente como: “difícil”, “muito difícil”, “fácil” e “muito fácil”. Tais valores estão contidos em um conjunto *fuzzy*. Cada conjunto *fuzzy*, A, é definido em termos de relevância a um conjunto universal, *U*, por uma função denominada de função de pertinência, associando a cada elemento *x* um número, $\mu A(x)$, no intervalo fechado [0,1] que caracteriza o grau de pertinência de *x* em A. O fator de pertinência pode assumir qualquer valor entre 0 e 1, representando completa exclusão e completa pertinência, respectivamente. A função de pertinência tem a forma $\mu A: X \rightarrow [0, 1]$.

Os sistemas *fuzzy* empregam um conjunto de regras do tipo “Se-Então” baseadas em variáveis linguísticas. Inicialmente, as variáveis de entrada passam por um processo denominado “fuzzificação”, onde é feito um mapeamento do conjunto de números reais para um conjunto *fuzzy*. Em seguida, efetua-se a inferência sobre o conjunto de regras fuzzy obtendo os valores dos termos das variáveis de saída. O mecanismo de inferência define a base para tomada de decisões. Por fim, as variáveis de saída passam por um processo denominado “defuzzificação” que consiste em transformar dados *fuzzy* para valores numéricos reais. Para isto, são utilizadas várias técnicas, tais como valor máximo, média dos máximos, média local dos máximos, centro de gravidade, ponto central da área e o centro da média [Driankov *et al*. 1993].

3. Trabalhos Relacionados

O desafio deste trabalho encontra-se no desenvolvimento de uma ferramenta que automatize o processo de avaliação de questões discursivas. Estudos neste sentindo ocorrem desde meados da década de 60, quando o sistema PEG (*Project Essay Grader*) foi desenvolvido para avaliar pequenas questões discursivas [Page 1967]. O PEG baseia-se principalmente na análise de modelo de características de superfície linguísticas de um bloco de texto. Assim, um ensaio é predominantemente classificado

ISSN: 2316-6541

com base em qualidade de escrita, não levando em conta o conteúdo. Sendo assim a correção de questões discursivas que por sua vez não existe limitação de escrita, seria ineficaz deste modo. Com o surgimento de novas técnicas como PLN (Processamento de Linguagem Natural) e EI (Extração da Informação) em meados da década de 90, retomaram-se as pesquisas e novas ferramentas foram desenvolvidas, a exemplos de IEA e E-Rater [Hearst 2000].

Outras pesquisas relacionadas à correção automática de testes discursivos vêm sendo realizadas de diversas formas, dentre elas podemos citar a proposta de Caldas e Favero (2009), que utiliza técnicas da Inteligência Artificial para realizar uma avaliação quantitativa e qualitativa sobre Mapas conceituais (MC). O trabalho é utilizado em ambientes (EAD) e a avaliação é realizada de duas maneiras: i) quantitativa através de um escore dado ao MC do estudante; e ii) qualitativa através de um "relatório-guia" fornecido ao estudante após o desenvolvimento de seu mapa conceitual. Segundo Novak (1984), mapas conceituais são representados de forma hierárquica representando, desta forma, a interdependência de conceitos. Entretanto, a metodologia de mapas conceituais força os alunos a seguirem uma estrutura que pode muitas vezes não captar o total conhecimento do aluno.

Os trabalhos citados anteriormente utilizam modelos estatísticos para classificação de textos. Nos modelos estatísticos quando a modelagem é definida, o modelo está pronto para ser testado e posteriormente usado. Porém, após ser definido o modelo sua mutação acarreta complicações que são de difícil manutenção para humanos. [Maia 2008] Sendo assim o uso da lógica *fuzzy* pode apresentar benefícios na classificação de textos, uma vez que sua metodologia se aplica na extração de conhecimento de especialistas de tal forma que seja fácil o entendimento para humanos.

4. SCATeDi - Sistema de Correção Automática de Testes Discursivos

Para o desenvolvimento da ferramenta apresentada neste trabalho, SCATeDi, utilizou-se as seguintes tecnologias: i) linguagem JAVA e a API para desenvolvimento de sistemas *fuzzy* em Java, *jFuzzyLogic*; e ii) o apoio computacional Ogma [Maia 2008], que permite a extração de termos das respostas dos alunos e do professor, desconsiderando *stopwords*[1].

4.1. Arquitetura da Ferramenta

A Figura 1 apresenta a arquitetura e o funcionamento do sistema SCATeDi. A etapa 1) compreende a interação entre o usuário e o sistema. O professor insere as respostas das questões dos alunos, bem como os gabaritos das questões. Uma vez feita à inserção das questões e das respostas, o sistema passa para etapa 2), na qual, por meio da ferramenta Ogma serão obtidos os termos sem a presença de *stopwords*. Na etapa 3) os termos do aluno são confrontados a um dicionário de sinônimos, onde o sistema verifica se existe algum sinônimo dos termos do aluno que seja correspondente a o termo do professor. Após a identificação dos sinônimos, o sistema continua e na etapa 4). é realizado a comparação entre os termos aluno e do professor. É importante salientar que para comparação entre os termos dos dois conjuntos são considerados os sinônimos e as variações de gênero e número das palavras, por exemplo, "realizado" e "realizadas" são

[1] Palavras que não são úteis para a recuperação de informações (exemplo palavras comuns, preposição, artigos, etc.)

ISSN: 2316-6541

consideradas palavras iguais. O resultado da comparação gera valores que serão atribuídos às variáveis linguísticas do sistema *fuzzy* na etapa 5). Na etapa 6) o sistema fuzzy repassa as variáveis para o processo de inferência onde são aplicadas as regras definidas por especialistas. O processo de defuzzificação é alcançado nas etapas 7) e 8), nas quais é possível obter um valor real do sistema e na etapa 9) o sistema fuzzy apresenta a nota ao usuário.

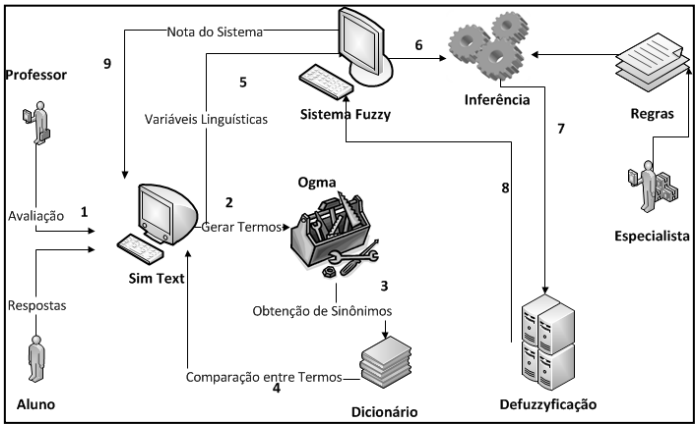

Figura 1. Arquitetura e Funcionamento do Sistema SCATeDi

Para calcular a nota do aluno, são utilizadas duas métricas: **cobertura** e **precisão**. **Cobertura** é o resultado da divisão da quantidade de termos do aluno iguais aos do professor pela quantidade total de termos do professor. **Precisão** é o resultado da divisão da quantidade de termos do aluno iguais aos do professor pela quantidade total de termos do aluno. Além da cobertura e precisão de cada aluno, SCATeDi gera também a cobertura global e precisão global da turma. Essas duas variáveis são responsáveis por armazenar a média da cobertura e da precisão de todos os alunos de uma determinada turma. Essas variáveis tem um importante papel para correção de questões discursivas, pois medem o desempenho da turma o que pode influenciar no cálculo da nota de cada aluno.

4.2. Modelagem *Fuzzy*

Para a entrada do sistema *fuzzy* foram definidas quatro variáveis linguísticas juntamente com seus respectivos termos linguísticos, são elas: “Precisão Individual”, “Cobertura Individual”, “Precisão Global” e “Cobertura Global”. Todas estas variáveis apresentam os termos linguísticos “Baixa”, “Média” e “Alta”. A variável “Cobertura Individual” representa a proporção de palavras-chave equivalentes às do professor e pode assumir valores entre 0 (zero) e 100 (cem). A “Precisão Individual” consiste na proporção de palavras-chave da resposta do aluno que não pertencem à resposta do professor (falsos positivos) e também assume valores entre 0 e 100. As variáveis “Precisão Global” e “Cobertura Global” correspondem, respectivamente, à média da precisão e da cobertura de todos os alunos cujas respostas foram inseridas no sistema. A saída do sistema é composta pela variável “Similaridade”, que pode receber valores do intervalo [0, 10]. A variável similaridade é composta de cinco conjuntos *fuzzy*, são eles “Muito Baixa”, “Baixa”, “Média”, “Alta” e “Muito Alta”.

A Figura 2 apresentam os gráficos relacionados às variáveis “Precisão Global” e “Similaridade”. Os gráficos das demais variáveis de entrada do sistema *fuzzy* foram omitidos por serem similares ao gráfico da variável precisão global.

Durante o desenvolvimento deste sistema *fuzzy* foram criadas onze regras orientadas por especialistas (professor do ensino superior e médio), que são apresentadas na Tabela 1.

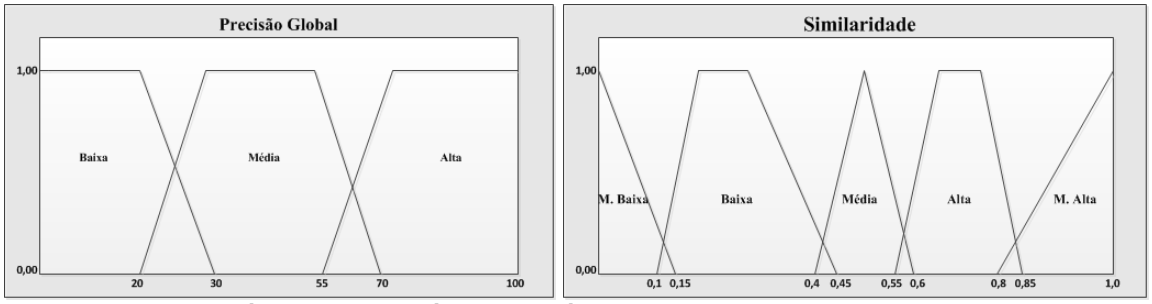

Figura 2. Gráficos das variáveis linguísticas precisão global e similaridade

Tabela 1. Regras do Sistema SCATeDi

#	Regras
R0	SE **cobertura individual** é Baixa a **similaridade** será Muito Baixa;
R1	SE **cobertura individual** é Media E **precisão individual** é Baixa a **similaridade** será Baixa;
R2	SE **cobertura individual** é Media E **precisão individual** é Media E **cobertura global** é Baixa E **precisão global** é Baixa a **similaridade** será Media;
R3	SE **cobertura individual** é Media E **precisão individual** é Media E **cobertura global** é Baixa E **precisão global** é Media a **similaridade** será Media;
R4	SE **cobertura individual** é Media E **precisão individual** é Media E **cobertura global** é Baixa E **precisão global** é Alta a **similaridade** será Media;
R5	SE **cobertura individual** é Media E **precisão individual** é Media E **cobertura global** não é Baixa a **similaridade** será Media;
R6	SE **cobertura individual** é Media E **precisão individual** é Alta E **cobertura global** é Baixa E **precisão global** é Baixa a **similaridade** será Alta;
R7	SE **cobertura individual** é Media E **precisão individual** é Alta E **cobertura global** não é Baixa a **similaridade** será Media;
R8	SE **cobertura individual** é Alta E **precisão individual** é Baixa E **cobertura global** é Alta E **precisão global** é Baixa a **similaridade** será Alta;
R9	SE **cobertura individual** é Alta E **precisão individual** é Baixa E **cobertura global** é Baixa E **precisão global** não é Baixa a **similaridade** será Media;
R10	SE **cobertura individual** é Alta E **precisão individual** não é Baixa a **similaridade** será Muito Alta;

As principais regras do sistema são comentadas demonstrando o papel que as mesmas desempenham no conjunto *fuzzy*. A regra R6 apresenta como o sistema *fuzzy* lida com o cenário em que a maioria dos alunos de uma turma obtiveram resultados insatisfatórios. Segundo esta regra, se a cobertura da resposta do aluno for MÉDIA, isto é, aproximadamente metade dos termos da resposta do aluno condizem com o gabarito fornecido do professor, e a precisão foi ALTA, o que significa que o aluno foi objetivo em sua resposta e não adicionou em sua resposta muitos termos que não dizem respeito ao que o professor esperava. Porém, a cobertura e a precisão globais revelam que o desempenho médio da turma em que este aluno está inserido foi BAIXO, ou seja, a maioria de seus colegas obtiveram resultados insatisfatórios. Segundo os especialistas consultados, este é um caso em que apesar da cobertura da resposta do aluno ter sido média, sua nota (variável similaridade) é considerada ALTA. A regra R9, por sua vez, retrata a situação na qual o aluno preencheu a avaliação com respostas corretas, porém contendo muitos falsos positivos (cobertura individual ALTA e precisão individual BAIXA), o que compromete a nota do aluno, sendo esta classificada como MÉDIA.

A Tabela 2 exemplifica a execução do sistema SCATeDi, apresentando as resposta de três alunos distintos para a seguinte questão: “Quais os fatores que contribuíram para a concentração do desenvolvimento industrial na região sudeste?”. A segunda coluna da Tabela 2 apresenta a resposta dada pelo professor como referência

para avaliação (linha 1), juntamente com as respectivas respostas dos alunos (linhas 2 à 4). As palavras em negrito representam os sinônimos, os quais foram considerados como termos equivalentes no cálculo das notas dos alunos. A terceira e quarta coluna mostram a precisão e a cobertura individuais obtidas pelos alunos. A quinta coluna, NS, apresenta a nota dada pelo sistema após passar por todas as etapas descritas na Figura 1. A sexta e ultima coluna, NP, descreve a nota dada pelo professor, após corrigir as respostas dos alunos com relação à questão mencionada anteriormente. A última linha apresenta a cobertura e precisão globais para esta questão.

Tabela 2. Exemplo de aplicação do sistema SCATeDi

Gabarito	A numerosa mão de obra, infraestrutura básica de transportes, **amplo** mercado consumidor, importantes fontes de recursos naturais, grande potencial hidrelétrico.	Precisão	Cobertura	NS	NP
Resposta Aluno 1	A existência de numerosa mão de obra, infraestrutura básica de transportes, **vasto** mercado consumidor, a existência de importação fontes de recursos naturais, grande potencial hidrelétrico.	0,83	0,93	9,49	10,00
Resposta Aluno 2	A **abundante** mão de obra por parte dos imigrantes europeus, a presença de uma infraestrutura de transportes formada por ferrovias.	0,42	0,31	5,00	4,00
Resposta Aluno 3	Vários fatores contribuem para desenvolvimento industrial, destacam-se numerosa mão de obra, infraestrutura básica de transportes, amplo mercado consumidor, importantes fontes de recursos naturais, **amplo** potencial hidrelétrico.	0,89	1,00	9,49	10,00
Cobertura Global: 0,75		**Precisão Global**: 0,71			

5. Avaliação do Sistema SCATeDi

Para verificar a aplicabilidade do sistema SCATeDi, realizou-se um estudo experimental com avaliações de três alunos da disciplina de geografia do programa Educação para Jovens e Adultos (EJA). Empregou-se três provas de alunos diferentes, cada uma contendo cinco questões, as quais foram corrigidas pelo professor antes da aplicação do sistema.

a) Definição do Estudo Experimental. O objetivo deste estudo é verificar se as notas obtidas com a correção de questões discursivas de avaliações de estudantes, quando o sistema especialista SCATeDi é utilizado (correção automática), equivalem às notas obtidas dadas por um professor (correção manual).

b) Seleção do Contexto. O estudo foi realizado com avaliações de alunos do segundo grau do programa EJA, no contexto da disciplina de Geografia e esse estudo não teve impacto na nota do aluno.

c) Formulação das Hipóteses. Foram elaboradas duas hipóteses para este estudo, as quais são apresentadas abaixo.

H_0: Não há diferença entre as médias das notas de um aluno, quando gerada pelo sistema SCATeDi e quando gerada manualmente pelo professor da disciplina ($NOTA_{SCATeDi} = NOTA_{Professor}$)
H_1: Há diferença entre as médias das notas de um aluno, quando gerada pelo sistema SCATeDi e quando gerada manualmente pelo professor da disciplina ($NOTA_{SCATeDi} \neq NOTA_{Professor}$).

d) Seleção das Variáveis. Variáveis independentes são aquelas manipuladas e controladas durante o estudo. Neste estudo, as variáveis independentes são as técnicas de correção de questões discursivas: Manual (realizada pelo professor) e Automática (realizada com a ferramenta SCATeDi). As variáveis dependentes são aquelas sob análise e cujas variações, com base nas mudanças feitas nas variáveis independentes, devem ser observadas. Neste experimento as notas das avaliação dos estudantes são consideradas como variáveis dependentes.

ISSN: 2316-6541

e) Seleção das Participantes. Os participantes do experimento foram selecionados por meio de amostragem não probabilística por conveniência.

f) Ameaças à Validade do Estudo Experimental. Um ponto que pode ter influenciado os resultados foi utilizar apenas de estudantes e professores de EJA como participantes do estudo. Contudo, não foram demonstradas expectativas a favor ou contra algumas das técnicas analisadas, para que os alunos e professores não fossem influenciados. Outro fator importante é a pequena quantidade de amostras utilizadas no estudo. Contudo, pretende-se replicar este experimento com maior número de participantes, bem como em outros contexto (por exemplo, outras disciplinas) a fim de tornar os resultados mais estatisticamente significativos.

g) Análise dos Dados Coletados. A Tabela 3 apresenta os dados coletados neste estudo experimental após a utilização do sistema SCATeDi. As colunas um e dois desta tabela, contêm os códigos dos alunos e das questões da avaliação utilizada neste estudo. A coluna três apresenta a nota gerada pelo sistema para cada questão da avaliação. A média das notas destas questões dá origem à nota total da avaliação, que é apresentada em linhas destacadas em azul. A coluna 4 mostra a nota fornecida pelo professor. É importante salientar que estas notas foram obtidas antes da execução do sistema SCATeDi, para evitar que os professores não fossem influenciados pelas notas do sistema. Por fim, a coluna 5 apresenta o erro gerado em cada questão. Este erro é obtido pela subtração da nota do sistema pela nota do professor, demonstrando a média de erros por prova.

Tabela 3. Dados do Estudo Experimental

Aluno	Questão	NS	NP	Erro (NS-NP)
A1	Q1	9,49	10,0	-0,51
	Q2	9,49	10,0	-0,51
	Q3	9,49	10,0	-0,51
	Q4	9,49	10,0	-0,51
	Q5	9,49	5,0	4,49
Média da Prova		**9,49**	**9,0**	**0,49**
A2	Q1	9,49	10,0	-0,51
	Q2	0,49	3,0	-2,51
	Q3	5,0	4,0	1,0
	Q4	5,0	10,0	5,0
	Q5	9,49	5,0	4,49
Média da Prova		**5,894**	**6,4**	**0,506**

Aluno	Questão	NS	NP	Erro (NS-NP)
A3	Q1	9,49	10,0	-0,51
	Q2	0,49	0,0	0,49
	Q3	9,49	10,0	-0,49
	Q4	0,49	0,0	0,49
	Q5	4,66	5,0	0,34
Média da Prova		**4,924**	**5,0**	**0.076**

É possível observar a partir dos discriminados na Tabela 3 que a nota dada aos estudantes pelo professor e a nota gerada pelo sistema se aproximaram. Em alguns casos, o erro foi pequeno, como na nota da questão 5 do estudante A3, cujo erro foi 0,34. Em outro caso, porém o erro foi excessivo, como no caso da questão 5 do aluno A2, cujo erro foi de 4,49. Contudo, ao se analisar as notas da avaliação como um todo, os erros podem ser considerados aceitáveis, uma vez que os mesmos variam de 0,076 até 0,506. A Figura 3 apresenta o gráfico com a comparação entre as notas dadas pelo professor e as notas geradas pelo sistema. No eixo X estão às quinze questões dos três estudantes. No eixo Y estão as notas que notas que variam de zero à dez. A linha em cor cinza escuro representa as notas geradas pelo sistema SCATeDi e a linha em cor cinza claro são as notas estabelecidas pelo professor da disciplina.

ISSN: 2316-6541

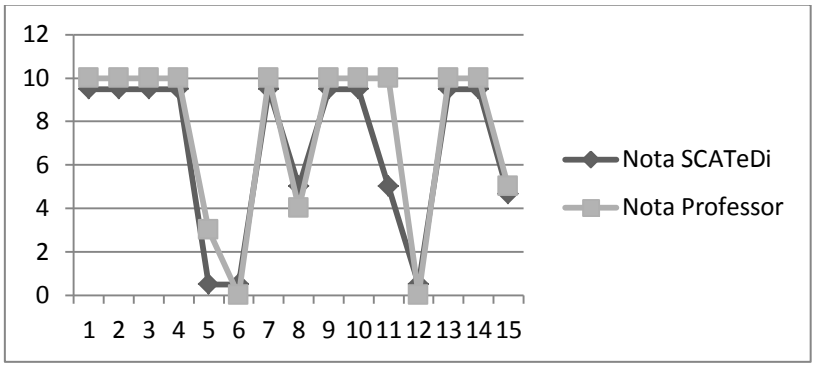

Figura 3. Comparação entre as notas do SCATeDi e as notas do professor

h) Teste das Hipóteses. Para verificar com maior confiabilidade o resultado apresentado no gráfico da Figura 3, realizou-se o teste das hipóteses deste estudo experimental. O objetivo deste passo é verificar, com algum grau de significância, se é possível rejeitar a hipótese nula (H_0) em favor de alguma da hipótese alternativa (H_1) com base no conjunto de dados obtido. Caso a hipótese nula seja refutada, prova-se que as notas médias dadas pelo professor e pelo sistema são diferentes. Caso contrário, é possível afirmar que as notas podem ser consideradas iguais com um determinado grau de significância.

A princípio, analisando os dados apresentados na Tabela 3, aparentemente o uso de correção automática das avaliações dos estudantes com a ferramenta SCATeDi não provocou modificações drásticas na média das notas dos alunos. Para comprovar esse efeito de forma estatística, aplicou-se o teste chamado *t-test* [Montgomery 2008]. Esse teste paramétrico é usado para comparar duas amostras independentes e checar se as médias de seus dados são estatisticamente equivalentes. No caso do estudo realizado, as amostras foram constituídas pelas notas dos alunos obtidas pelo sistema SCATeDi e pelo professor.

Uma vez que foram analisadas três provas de três alunos diferentes, a aplicação do *t-test* ao conjunto amostral de dados foi realizada em três etapas. Na primeira etapa, compararam-se as amostras relativas às notas das questões da prova do aluno A1 (Tabela 3). Já na segunda etapa, a comparação foi feita com as notas das questões do aluno A2. Por fim, na última etapa, analisou-se as notas do aluno A3. Para os propósitos deste estudo, em cada etapa do teste utilizou-se o menor grau de significância α com o qual fosse possível rejeitar a hipótese nula, assumindo-se um grau de significância máximo de 5%. A análise dos dados deste estudo foi realizada utilizando um pacote de extensão estatístico para o software Excel denominado Analyse-it[2].

Primeira Etapa: Baseada nas duas amostras independentes da primeira etapa do experimento: $NOTA_{SCATeDi}$ = {9,49; 9,49; 9,49; 9,49; 9,49}, média = 9,49; e $NOTA_{Professor}$ = {10,00; 10,00; 10,00; 10,00; 5,00}, média = 9,00. **Segunda Etapa**: baseada nas duas amostras independentes da segunda etapa do experimento: $NOTA_{SCATeDi}$ = {9,49; 0,49; 5,0; 5,0; 9,49}, media = 5,894; e $NOTA_{Professor}$ = {10,0; 3,0; 4,0; 10,0; 5,0}, media = 6,4. **Terceira Etapa**: baseada nas duas amostras independentes da terceira fase do experimento: $NOTA_{SCATeDi}$ = {9,49; 0,49; 9,49; 0,49; 4,66}, media n=4,924 ; e $NOTA_{Professor}$ = {10,0; 0,0; 10,0; 0,0; 5,0}, média = 5,0. Em

[2] www.analyse-it.com/

todas as três etapas, não foi possível rejeitar a hipótese nula H_0, concluindo-se que as médias das notas dos alunos A1, A2 e A3 são iguais em um nível de significância de 5%, ou seja, pode-se afirmar com 95% de certeza que as médias das notas destes alunos são iguais, tanto para as notas geradas pelo sistema SCATeDi, quanto que para as notas atribuídas pelo professor.

6. Considerações Finais e Trabalhos Futuros

Dentre os tipos de Avaliação, a avaliação discursiva pode apresentar os melhores resultados, pois por meio dela o discente tem a capacidade de expressar seu conhecimento de forma ímpar em relação aos demais testes (objetivos e orais). No entanto, o processo de correção de provas discursivas exige do educador esforço considerável.

O presente trabalho propõe a utilização de um sistema inteligente para avaliação automática de questões discursivas. Por meio de um estudo experimental realizado e descrito neste artigo, há indícios de que a ferramenta proposta seja aplicável e que apresente acurácia satisfatória. Como trabalhos futuros, pretende-se: i) desenvolver um módulo para correção automática de questões discursivas em ambientes de EAD, como o Moodle, por exemplo; ii) realizar outros estudos experimentais, com amostras maiores e em outros contexto (por exemplo, em disciplinas de cursos de graduação em ciência da computação) para verificar a aplicabilidade e acurácia da ferramenta proposta com maior significância estatística.

Referências Bibliográficas

Abrão, I. C. e Rayel, F. and Abrão, M. A. V. L. (2004), QUESTCOMP: Ferramenta para Avaliação de Aprendizado à Distância. WCETE, Guarujá.

Bezerra, M. A. (2008), Questões discursivas para avaliação escolar v. 30, n. 2.

Caldas, V. M. and Favero, E. L. (2009), Uma Ferramenta de Avaliação Automática para Mapas Conceituais como Auxílio ao Ensino em Ambientes de Educação a Distância.

Driankov, D. (1993), Hellendoorn, H.; Reinfrank, M. "An Introduction to Fuzzy Control". Springer-Verlag.

Hearst, M. (2000), The debate on automated essay grading. *IEEE Intelligent Systems, 15*(5), 22-37, IEEE CS Press.

Maia, L. C. G. (2008), Uso de sintagmas nominais na classificação automática de documentos eletrônicos. 2008. Tese (Doutorado em Ciência da Informação) - Universidade Federal de Minas Gerais - UFMG. Belo Horizonte, 2008.

Montgomery, D. C. (2000), Design and Analysis of Experiments, 5 ed., Wiley. Souza, O. T. L. (2004), Desenvolvimento de um Modelo Fuzzy para Determinação do Latente com Aplicação em Sistemas de Irrigação. Dissertação de Mestrado, UNESP.

Moodle. Disponível em: http://moodle.org/. Acessado em: Julho de 2012.

Novak, J.D. and Gowin D.B. (1984), Learning how to learn. New York and Cambridge, UK: Cambridge University Press.

Nozawa, E. H. and Oliveira, E. H. T. (2006), Simulador e-JLPT: Um Software de Apoio Educacional com Enfoque em Hipermídia Adaptativa. XVII SBIE. DF.

Page, E.B. (1967). Grading essays by computer: Progress report. Proceedings of the 1966 Invitational Conference on Testing (pp. 87-100). Princeton.

Perrenoud, P. (1999) Avaliação: da excelência à regulação da aprendizagem – entre duas lógicas, Porto Alegre: Artes Médicas Sul.

Zadeh, L. A. (1965), Fuzzy Sets. *Information and Control*, v. 8, p. 338-353.

ISSN: 2316-6541