

RadioAssess: um sistema para correção automática de provas dissertativas em Radiologia

**Pedro H. Candido¹, Rafael Lamare², Ivan T. Pisa², Henrique M. Lederman²,
Fátima L. S. Nunes¹**

¹Laboratório de Aplicações de Informática em Saúde (LApIS)
Escola de Artes, Ciências e Humanidades (EACH)
Universidade de São Paulo (USP) - São Paulo - SP - Brasil

²Laboratório de Mineração de Dados em Saúde (Saúde 360)
Escola Paulista de Medicina (EPM)
Universidade Federal de São Paulo (UNIFESP)

{pedro.candido,fatima.nunes}@usp.br, {rafael.lamare,ivanpisa,
henrique.lederman}@gmail.com

Abstract. *Clinical cases records are valuable subsidies for medical education in the Radiology field. However, correcting tests that use essay questions are still a challenge mainly because the variety of terms used in this area. This work presents the RadioAssess system aiming at supporting the correction of essay questions in Radiology, by using pre-processing of text and two correction approaches (with and without synonyms). The results indicated that the corrections are consistent with the manual correction of the teacher, according to the thesaurus provided. The system makes correction in a fast and flexible way and allows reuse synonyms recorded in previous tests.*

Resumo. *O registro de casos clínicos é um valioso subsídio para a educação na área de Radiologia. No entanto, a correção de avaliações usando questões dissertativas se mostra complexa devido principalmente à variedade de termos empregados na área. Este trabalho apresenta o sistema RadioAssess como apoio à correção de questões dissertativas em Radiologia, usando pré-processamento de texto e duas abordagens de correção (com e sem sinônimos). Os resultados indicaram que as correções são consistentes com a correção manual do professor, de acordo com o dicionário de sinônimos fornecido. O sistema efetua a correção com rapidez e flexibilidade e permite reusar sinônimos empregados em avaliações anteriores.*

1. Introdução

O registro de casos clínicos é de extrema importância na Medicina para que a qualidade nos serviços seja garantida, assim como para disponibilizar históricos e viabilizar estudos posteriores. Os históricos são disponibilizados tanto para fins acadêmicos favorecendo o aprendizado de estudantes, quanto para fins jurídicos em defesa do médico se questões sobre sua conduta e escolhas forem levantadas [Arnold *et al.* 2009, Mishra *et al.* 2009]. Por isso, existe a necessidade de descrever de forma coerente os processos, sinais e sintomas que conduziram ao diagnóstico dado, tendo em vista a saúde do paciente [Hauben *et al.* 2005].

A capacidade de formar diagnósticos corretos tem suas bases na educação médica e na forma como os estudantes de Medicina são avaliados. Como Buriasco (2000) observa, a avaliação está presente nas diversas áreas da atividade humana. De acordo com o mesmo autor, qualquer forma de avaliação pressupõe um julgamento com base em uma concepção implícita ou explícita.

A resolução de casos clínicos é uma tarefa comum em várias disciplinas acadêmicas da área de saúde, tanto para o treinamento do aluno, quanto para realizar avaliações de seu aproveitamento durante o curso [Grossman e Cardoso 2006].

Dado este cenário, recursos que facilitem a avaliação e o treinamento de profissionais de saúde podem contribuir para o desenvolvimento da sua formação, colaborando, consequentemente, para a qualidade do atendimento prestado ao paciente. Particularmente, a correção de provas dissertativas na área de Radiologia tem sido um desafio a ser superado na educação médica.

Experimentos conduzidos com estudantes da área resultaram em bom desempenho dos mesmos (obtenção de notas altas) em avaliações que consideraram unicamente questões objetivas. No entanto, os mesmos estudantes que obtiveram bom desempenho em provas objetivas não foram capazes de identificar achados importantes em imagens médicas quando empregadas questões dissertativas. Paradoxalmente, tem havido um crescente emprego de questões objetivas nas avaliações devido à dificuldade de correção de questões dissertativas. Trabalhos de correção de questões discursivas usam em geral a abordagem de comparação n-grama, e trazem resultados satisfatórios, em torno de 80% de assertividade [Santos *et al.* 2007]. Outros trabalhos tendem a usar processamento de linguagem natural, para, por exemplo, gerar Unidades de Leitura a partir de regras livres de contexto [Salton *et al.* 2013].

Apesar de existirem trabalhos similares de correção automatizada em outras áreas, a correção de provas dissertativas na área de Radiologia ainda constitui um desafio a ser vencido, principalmente porque a estruturação da prova é diferenciada em relação a outras áreas (obrigatoriamente são utilizadas imagens médicas e a estrutura da prova segue um padrão de diagnóstico), a correção deve ser rápida e devem ser observados termos correlatos.

O objetivo deste artigo é apresentar um sistema de informação denominado *RadioAssess* desenvolvido para propiciar correção de questões dissertativas de avaliações na área de Radiologia. O algoritmo de correção compara termos presentes no gabarito do professor e as respostas do aluno com abordagens usando os textos originais

e textos pré-processados. Como alternativa foi também empregada uma abordagem com o uso de sinônimos.

Para atingir o objetivo apresentado, este artigo está estruturado da seguinte forma: na Seção 2 são citados trabalhos correlatos que abordam a correção de questões e o pré-processamento de textos; na Seção 3 é descrita a metodologia utilizada, apresentando as abordagens de pré-processamento e correção utilizadas e detalhando as técnicas de implementação e validação; na Seção 4 são apresentados os resultados, incluindo um estudo de caso no qual são feitas comparações entre o método automático e o manual. Por fim, na Seção 5 são apresentadas as conclusões embasadas nos resultados obtidos.

2. Trabalhos correlatos

Apesar de existirem trabalhos na literatura atual que abordam a questão de correção automática de questões dissertativas, em geral a área de educação médica é pouco explorada, conforme apresentado a seguir.

Os principais casos de sucesso na correção automatizada de questões estão descritos nos trabalhos de revisão de [Valenti e Cucchiarelli, 2003] e [Dikli, 2006]. Neles é possível verificar as diferentes abordagens adotadas e o desempenho associado a cada algoritmo. Todavia, constata-se a ausência de um sistema de correção especializado na área médica.

As técnicas de mineração de texto, em especial as de pré-processamento textual como lematização, remoção de palavras repetidas e *stop words*, são especialmente úteis ao presente problema. Algoritmos avançados de *string matching* podem ser usados para tolerar pequenas falhas de escrita ou erros de digitação no momento da correção. Ademais, técnicas de processamento de linguagem natural e aprendizado de máquina podem ser usadas na evolução e construção de sistemas corretores mais robustos.

Santos et al. (2007) apresentam uma proposta de avaliação automática de questões discursivas pequenas (de 15 a 30 palavras) com o uso de n-gramas para medir a similaridade entre textos. De acordo com os autores em questões conceituais a abordagem atingiu acurácia média de mais de 80%. O algoritmo tem base no teste e escolha das melhores entre 11 técnicas de n-grama, sendo elas descritas e exemplificadas nas Tabelas 1 e 2.

Com base em regras sintáticas livres de contexto para correção de unidades de leitura, Salton et al. (2013) utilizam processamento de linguagem natural para o processo de reconhecimento e correção de respostas. A metodologia usada obtém as produções da gramática de cada resposta e, em seguida, processa a resposta em *tokens* (palavras), valida essas respostas e pontua cada resposta fazendo as devidas comparações. As linguagens ditas livres de contexto são utilizadas para representar características sintáticas de sentenças, e foram empregadas no trabalho citado para gerar tabelas de *parsing*, nas quais *tokens* retirados das respostas dos alunos foram comparados com palavras que são esperadas em respostas corretas, categorizadas por importância. Os autores afirmam que as respostas corrigidas pela ferramenta apresentaram afinidade acima de 90% em comparação com a correção realizada por especialistas humanos para as respostas do grupo de controle do estudo de caso dessa técnica.

Tabela 1. Uni-grama, bi-grama e tri-grama de caracteres de uma palavra, com a presença e ausência de espaços em branco.

Técnica	Palavra: "Árvores"
1. Uni-gramas	[á,r,v,o,r,e,s]
2. Bi-gramas com brancos	[(∅, á), (á, r), (r, v), (v, o), (o, r), (r, e), (e, s), (s, ∅)]
3. Tri-gramas com brancos	[(∅, ∅, á), (∅, á, r), (á, r, v), (r, v, o), (v, o, r), (o, r, e), (r, e, s), (e, s, ∅), (s, ∅, ∅)]
4. Bi-gramas sem brancos	[(á, r), (r, v), (v, o), (o, r), (r, e), (e, s)]
5. Tri-gramas sem brancos	[(á, r, v), (r, v, o), (v, o, r), (o, r, e), (r, e, s)]

Tabela 2. Uni-grama, bi-grama e tri-grama de palavras de uma frase, com remoção de *stop words* e com o uso da técnica de *stemming*.

Técnica	Frase: "As árvores altas que cortam os céus"
6. Uni-gramas	[as, árvores, altas, que, cortam, os, céus]
7. Bi-gramas	[(as, árvores), (árvores, altas), (altas, que), (que, cortam), (cortam, os), (os, céus)]
8. Tri-gramas	[(as, árvores, altas), (árvores, altas, que), (altas, que, cortam), (que, cortam, os), (cortam, os, céus)]
9. Uni-gramas, sem <i>stop words</i> , com <i>stemming</i>	[arvor, alt, cort, ceu]
10. Bi-gramas, sem <i>stop words</i> , com <i>stemming</i>	[(arvor, alt), (alt, cort), (cort, ceu)]
11. Tri-gramas, sem <i>stop words</i> , com <i>stemming</i>	[(arvor, alt, cort), (alt, cort, ceu)]

Em uma abordagem semelhante às já apresentadas, Kaur e Jyoti (2012) trazem uma abordagem de avaliação automatizada em respostas curtas de uma linha de texto livre em Ciência da Computação, empregando técnicas de pré-processamento das respostas, removendo *stop words*, corrigindo erros de digitação e gramática e fazendo o *stemming* das palavras. Sua metodologia para fazer a correção foi criar uma lista de respostas de alunos e outra com a resposta correta, remover toda a forma de pontuação, artigos, *stop words*, palavras duplicadas e fazendo a contagem de tamanho de palavras nessas duas listas. Após este pré-processamento, os erros gramaticais são corrigidos e as

palavras são comparadas buscando seu grau de relevância entre a resposta do aluno em relação à resposta tida como correta.

3. Metodologia

Tendo como objetivo principal o desenvolvimento de um sistema para correção automática de questões dissertativas de provas na área de Radiologia, este trabalho visa a disponibilizar uma ferramenta de utilização fácil e rápida, e com assertividade na correção das questões. Foi implementado um sistema web (disponível em <http://143.107.58.159/lapis/radioassess/>), o qual possui a estrutura para cadastro de avaliações, gabaritos e respostas, um *parser* de *strings* para processar o conteúdo inserido como gabarito, e um corretor integrado, que usa dois métodos de correção (com e sem sinônimos).

A fase de coleta dos dados utilizou o sistema web de apoio implementado para o cadastro das avaliações e o recebimento de respostas dos alunos. A correção das provas consistiu na execução de três passos: (1) cadastro da avaliação com questões e gabarito para cada questão; (2) coleta das respostas dos alunos e (3) comparação de cada resposta com o gabarito, usando uma abordagem sem sinônimos e outro com sinônimos. Esses passos são detalhados nas seções a seguir.

3.1 Estrutura da plataforma web

O sistema foi elaborado para funcionar no paradigma de computação em nuvem, ou seja, um servidor remoto (que poderá ser escalado para mais servidores) é responsável pelo armazenamento e processamento, não sendo necessária a instalação de um software específico, mas somente conexão com a Internet. O sistema segue a estrutura padrão de uma aplicação web, tendo como pontos principais o gerenciamento de usuários e o gerenciamento de avaliações.

A Figura 1 ilustra os três níveis de usuários considerados no sistema, sendo que o nível de Administrador tem a função de cadastro de Professores e Alunos, descritas no diagrama de caso de uso.

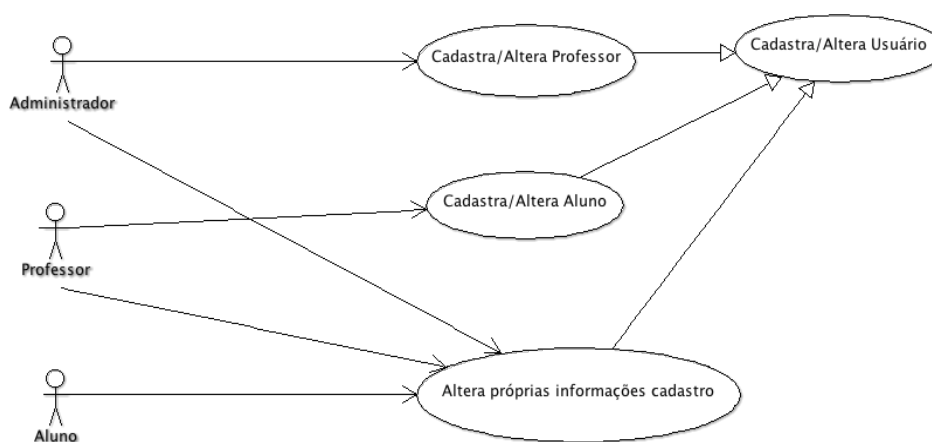


Figura 1. Diagrama de casos de uso do sistema *RadioAssess*.

Na Figura 2 é apresentado um diagrama de caso de uso, definindo-se as funções que os atores Professor e Aluno devolvem no sistema em relação às avaliações, desde o seu cadastro até a aplicação da prova, finalizando com a correção da mesma.

Todas as ferramentas utilizadas estão disponíveis sob alguma forma licença de software livre. O servidor web consiste de uma instalação do Apache (Apache Software Foundation, 2014) em uma máquina com sistema operacional Linux, executando a linguagem de programação PHP (PHP Group, 2014). O sistema foi escrito em linguagem de programação PHP utilizando o *framework* Symfony (Potencier, 2014). O Sistema Gerenciador de Banco de Dados utilizado foi o MySQL (Oracle Corporation, 2014). Para a camada de apresentação, que é renderizada no lado do cliente pelo navegador, foi usado HTML, CSS e JavaScript (World Wide Web Consortium, 2014).

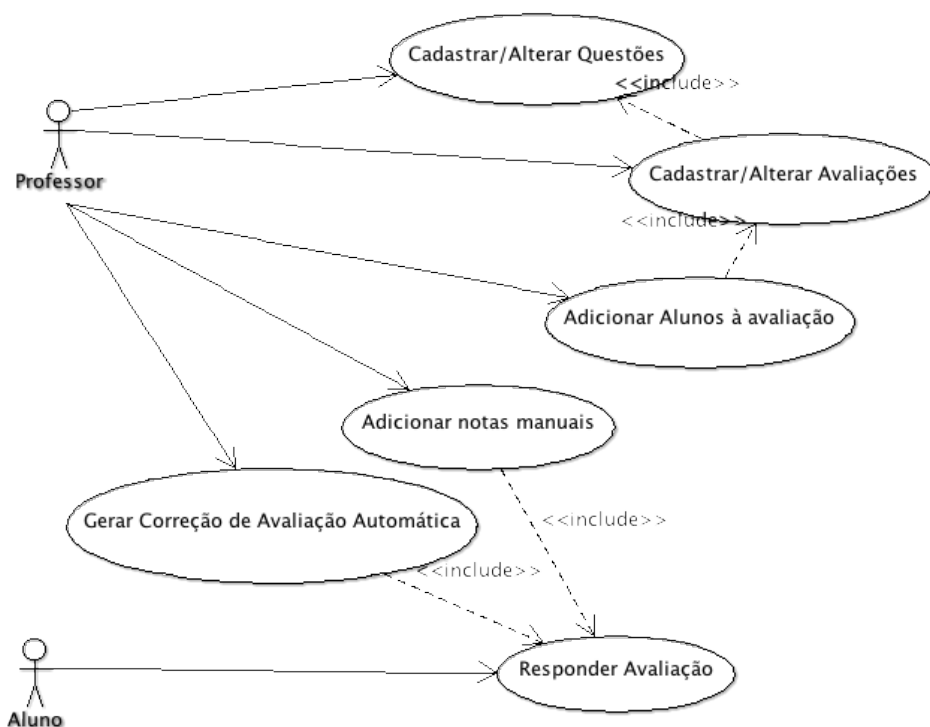


Figura 2. Gerenciamento de Avaliações

3.2 Modelagem de dados

Conforme apresentado no Diagrama Entidade-Relacionamento da Figura 3, a modelagem de dados considera tabelas de acesso e permissão, bem como as tabelas que são necessárias para a correção, como termos, avaliação, questões e respostas.

No modelo apresentado, o armazenamento de termos e seus sinônimos é a parte primordial para o funcionamento do presente sistema. A entidade "termo" armazena o texto do termo cadastrado pelo professor e a entidade "termo_sinonimo" armazena o relacionamento entre dois termos, indicando que são sinônimos. No diagrama, cada termo cadastrado está relacionado tanto à questão quanto a uma categoria, fazendo com que um glossário comece a ser formado, já que cada novo termo será atrelado à

categoria da questão. Esta modelagem permite que em usos futuros os termos sinônimos possam ser sugeridos ao professor durante a composição do gabarito, considerando usos anteriores do termo. Ao mesmo tempo, permite que sinônimos sejam considerados particularmente para cada questão.

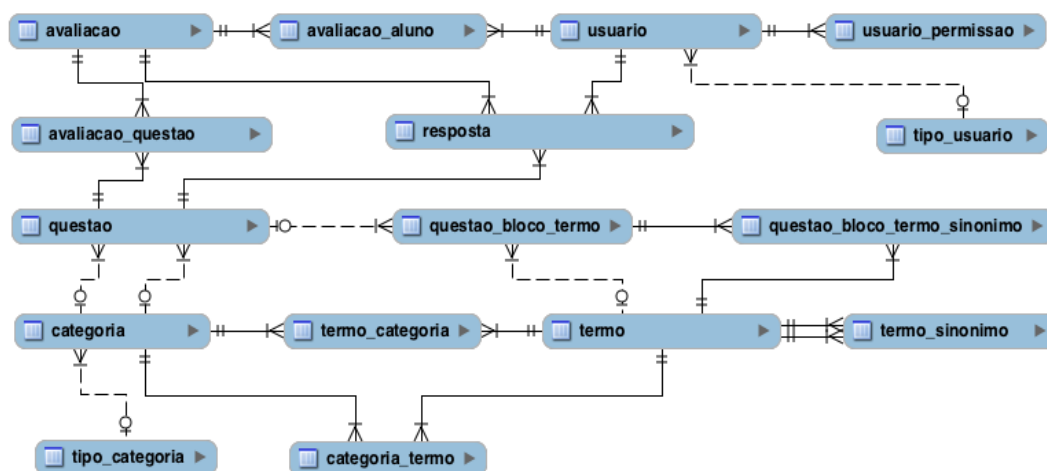


Figura 3. Modelo Entidade-Relacionamento do sistema *RadioAssess*.

3.3 Apresentação do sistema

O Sistema *RadioAssess* é composto por questões, avaliação e corretor, conforme apresentado nas seções a seguir.

3.3.1 Cadastro de questões

Para usar o sistema, inicialmente devem ser cadastradas as questões. Conforme mostrado na Figura 4, a interface de cadastro é dividida em três partes: Questão, Glossário e Imagens. É importante destacar que esta estrutura obedece um padrão na Radiologia, sendo que a presença de imagens é intrínseca à natureza da área.

A primeira divisão (Questão) apresenta as informações que trazem sua pertinência para o professor durante a elaboração da avaliação: título, a parte do corpo humano à qual a questão se refere e o tipo de exame que será abordado. A interface também disponibiliza o enunciado da questão que será apresentado para o aluno (Figura 4).

Questão

Título * <input type="text" value="Sinusopatia crônica. Controle coi"/>	Parte do Corpo * <input type="text" value="Cabeça"/>
Enunciado * <input type="text" value="Sinusopatia crônica. Controle com TC."/>	Tipo de Exame * <input type="text" value="Radiografia"/>

Figura 4. Cadastro de questões no sistema *RadioAssess*.

A segunda parte do cadastramento de questões (Glossário) oferece a possibilidade de gerenciar os termos que devem ser buscados nas respostas dos alunos (Figura 5). Esta parte deve ser obrigatoriamente preenchida pelo professor (no caso, médico radiologista), que é o ator que detém o conhecimento acerca do conteúdo.

Figura 5. Inserir termos no glossário no sistema *RadioAssess*.

No exemplo da Figura 5 são disponibilizadas três abas: "Achados das Imagens", "Diagnósticos Diferenciais" e "Diagnóstico Final". Cada uma dessas abas possibilita o cadastro de um ou mais termos e seus sinônimos em forma de *tokens* de texto. Esses termos são divididos em "certos" e "errados", a fim de que na correção sejam adicionados ou descontados pontos para compor a nota de cada resposta, dependendo do seu tipo e da configuração inserida no cadastro da avaliação.

A última parte do cadastro de questões ("Imagens") é responsável pela inserção de até três imagens, que são apresentadas ao aluno durante a resolução da avaliação e para os professores enquanto estiverem revendo as questões (Figuras 6 e 7).

O cadastro de Avaliação é efetuado ao inserir um nome para a avaliação e as suas questões. A Figura 7 apresenta os atributos que devem ser preenchidos para cada questão.

A Figura 7 mostra a interface do módulo de cadastramento de avaliação, com o exemplo de configuração de duas diferentes questões. Para adicionar novas questões, o Professor deve acionar o botão "Adicionar questão".

Adicionar Imagens




Imagem 1

Sem arquivo...

Escolher

Imagem 2

Sem arquivo...

Escolher

Imagem 3

Sem arquivo...

Escolher

Figura 6. Espaço para inserção de imagens, com uma imagem adicionada, no cadastro de questões no sistema RadioAssess.

Adicionar Questão

Questão: Sinusopatia crônica. Control...

Peso da Questão: 5

Pesos Respostas:

Certas: 10

Erradas: 10

Outras: 2

Pesos Blocos:

Achados das Imagens: 2

Diagnósticos Diferenciais: 2

Diagnóstico Final: 6

Questão: Pós-transplante - RX Face pe...

Peso da Questão: 5

Pesos Respostas:

Certas: 10

Erradas: 5

Outras: 2

Pesos Blocos:

Achados das Imagens: 10

Diagnósticos Diferenciais: 10

Diagnóstico Final: 10

Figura 7. Adição das questões em uma avaliação no sistema RadioAssess.

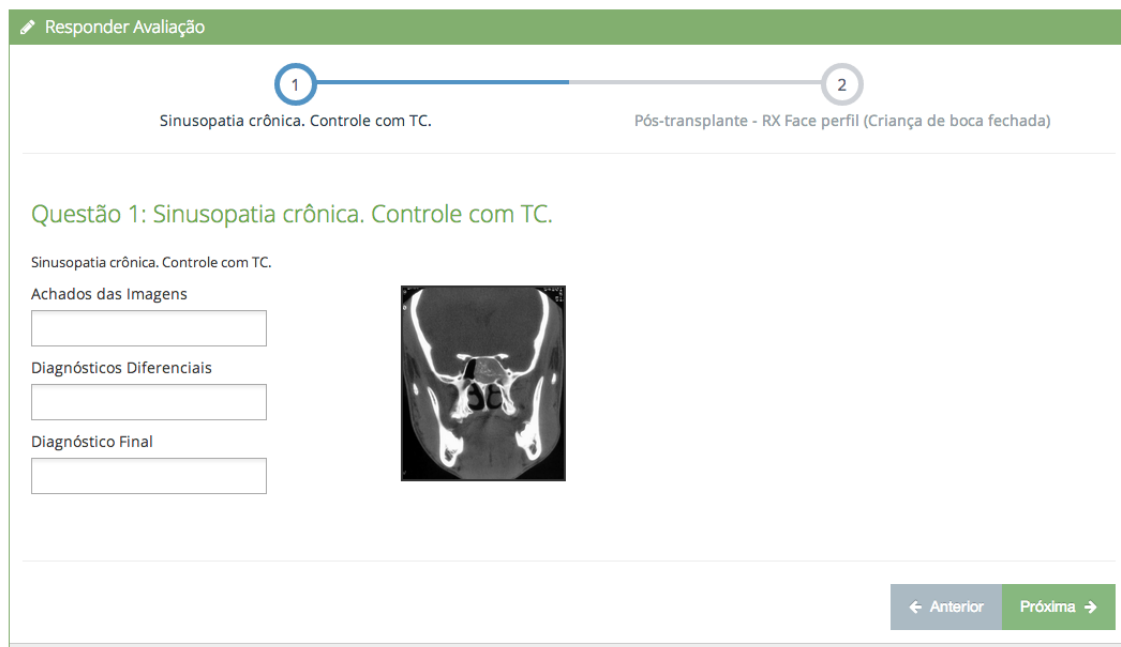
Os campos que atribuem pesos às questões têm a função de definir a porcentagem da nota da questão em relação à prova total, permitindo que o professor defina uma questão como mais importante que outra e atribua a ela pontuação diferenciada. Essa definição de pontuação é cadastrada para cada avaliação, podendo ser alterada em uma nova avaliação incluída posteriormente com as mesmas questões.

Os campos relacionados a pesos de respostas informam a pontuação que cada termo encontrado deve adicionar ou descontar da nota da questão durante o processo de correção de termos certos e errados. Os pesos dos blocos referem-se à participação que cada bloco ("Achados das Imagens", "Diagnósticos Diferenciais" e "Diagnóstico Final") apresenta na pontuação final da questão.

Por exemplo, para a primeira questão da Figura 7, se o aluno acertar completamente os "Achados das Imagens" e "Diagnósticos Diferenciais" e acertar apenas a metade do "Diagnóstico Final", será obtida a nota final 7. O mesmo desempenho do aluno para a segunda questão resultaria na nota 8,3.

3.3.3 Respostas de Avaliações

A resposta da avaliação de cada aluno é recebida em uma interface que permite a coleta dos dados em uma sequência de passos (Figura 8). Para cada questão são apresentados o título, o enunciado e as imagens inseridas.



Responder Avaliação

1 2

Sinusopatia crônica. Controle com TC. Pós-transplante - RX Face perfil (Criança de boca fechada)

Questão 1: Sinusopatia crônica. Controle com TC.

Sinusopatia crônica. Controle com TC.

Achados das Imagens

Diagnósticos Diferenciais

Diagnóstico Final

← Anterior Próxima →

Figura 8. Tela de respostas de avaliação no sistema RadioAssess.

3.4 Implementação do corretor

O algoritmo de correção empregado no presente trabalho é composto por duas partes principais: o pré-processamento do texto inserido pelo aluno e a comparação do texto com o gabarito do professor.

O pré-processamento de texto é feito de duas formas: redução de palavras para seus radicais (pseudo-lemas), chamado de *stemming* e remoção de *stop words* e caracteres especiais. *Stop words* são palavras que não agregam valor à resposta, como, por exemplo, preposições e artigos.

Para a redução de termos para radicais foi usado o algoritmo de Orengo e Huyck (2001), voltado para a língua portuguesa, no qual são definidas regras de limpeza da palavra para serem cumpridas em oito etapas, até ser encontrado o radical desta:

1. Redução do plural
2. Redução do feminino
3. Redução do aumentativo/diminutivo
4. Redução adverbial
5. Redução nominal
6. Redução verbal
7. Remoção de vogais
8. Remoção de acentuação

Para cada um desses passos o tamanho da palavra é verificado. Caso este tamanho seja maior do que o mínimo permitido, é verificado se a palavra não está

dentro das exceções e, caso não esteja, o sufixo é trocado ou removido. As regras de *stemming* para a língua portuguesa foram definidas por Orengo (2004).

Dessa forma, é necessário apenas estruturar as regras. A Figura 9 apresenta exemplos de algumas regras utilizando a estrutura YAML (Evans, 2014). Conforme explica Orengo (2004), a estrutura YAML ("*YAML Ain't Markup Language*") usada para representar as regras apresenta vantagens de uso por ser de fácil leitura por máquinas, assim como por pessoas, além de facilidade de implementação e possibilidade de troca de dados entre qualquer linguagem de programação.

```
plural:
- { suffix: ões, replacement: ão, size: 3, exceptions: { } }
- { suffix:ães, replacement: ão, size: 1, exceptions: [mães] }

feminino:
- { suffix: inha, replacement: inho, size: 3, exceptions: [rainha, linha, minha] }

aumentativo_diminutivo:
- { suffix: ão, replacement: '', size: 3, exceptions: [camarão, chimarrão, canção, coração] }
```

Figura 9. Exemplo de regras, escritas em YAML.

Um exemplo que poderia ser usado para exemplificar as regras citadas é a palavra "corações", a qual teria o sufixo "ões" substituído na validação de plural (passo 1) por "ão". Na sequência, ao aplicar o passo 3 (redução de aumentativo/ diminutivo), o sufixo "ão" permaneceria inalterado, visto tratar-se de uma exceção. O resultado, então, seria a palavra "coração". Palavras como "minha" não são reduzidas em regras como a do sufixo "inha" na regra de feminino, devido à determinação do tamanho mínimo da palavra e, por isso, não há necessidade de serem incluídas nas exceções.

Com as regras descritas, o algoritmo faz validações em sequência, respeitando a ordem descrita nas etapas acima. O processo de *stemming* ocorre em dois momentos distintos no sistema *RadioAssess*: o primeiro durante o cadastro de questões e o segundo durante a resposta dos alunos. Os radicais são armazenados junto com o texto original inserido pelos usuários, seja Professor ou Aluno.

Tendo esses dados já armazenados, o algoritmo faz a comparação dos termos inseridos nos cadastros anteriormente descritos e atribui uma nota para cada questão e, consequentemente, para a prova, de acordo com os critérios definidos pelo professor.

Três abordagens foram usadas para a correção de questões considerando-se o pré-processamento:

1. Texto original, abordagem na qual a resposta é verificada da forma que o aluno a inseriu, sem pré-processamento.
2. *stop words*, na qual a resposta tem todos os caracteres especiais processados e removidos (quando necessário) e *stop words* também sendo removidas.
3. *Stemming*, na qual as respostas resultantes da abordagem anterior (*stop words*) são ainda submetidas ao processo de *stemming* para redução de seus termos aos seus radicais.

A Tabela 3 apresenta exemplos de um texto usando essas três abordagens.

Adicionalmente, para cada uma dessas abordagens de pré-processamento, duas abordagens de correção foram empregadas: sem sinônimos e com sinônimos. As combinações das abordagens de pré-processamento com as abordagens de correção geraram seis conjuntos de resultados, apresentados na seção 4.

Tabela 3. Exemplos de resposta processada

Tipo de pré-processamento	Resposta
Texto original	Tomografia de crânio corte coronal. Seio esfenoidal esquerdo, aumento densidade partes moles, hipoaeração, calcificações de permeio. assimétrica, espessamento mucoso ocupando todo o seio com imagens de hiperatenuação que pode representar áreas de calcificação e hipoaeração.
<i>stop words</i>	tomografia cranio corte coronal seio esfenoidal esquerdo aumento densidade partes moles hipoaeracao calcificacoes permeio assimetrica espessamento mucoso ocupando seio imagens hiperatenuacao representar areas calcificacao hipoaeracao
<i>Stemming</i>	tomograf crani cort coron sei esfenoid esquerd aument densidad part mol hipoaeraca calcificaco permei assimetr espess muc ocup sei imag hiperatenuaca represent are calcificaca hipoaeraca

3.5 Estudos de caso

Para a comparação de resultados foram utilizados dois estudos de caso: uma avaliação de controle e uma avaliação com dados reais, com uma questão respondida por dez alunos. O objetivo da avaliação de controle foi comprovar que a implementação estava adequada aos requisitos definidos. O gabarito, as questões e as notas decorrentes da correção manual para o segundo estudo de caso foram disponibilizados por um professor radiologista experiente.

Os dados que compuseram as respostas para a avaliação de controle foram gerados considerando o gabarito da prova real fornecida, tendo sido criados quatro respostas:

1. resposta retirada do próprio gabarito do professor, que deveria receber nota máxima.
2. mesma resposta retirada do gabarito do professor, mas sendo removida a segunda metade dela, sendo esperando que o sistema atribuisse metade da nota máxima.
3. resposta completamente incorreta.
4. resposta em branco.

A questão proposta pelo professor foi: "Quais são os achados da imagem anexa?", sendo que a imagem citada está apresentada na Figura 10.

O gabarito está apresentado na Tabela 4. As respostas da avaliação de controle e as respostas reais avaliadas são apresentadas, respectivamente, nas Tabelas 5 e 6.



Figura 10. Imagem usada na avaliação diagnóstica pelos alunos.

Tabela 4. Gabarito disponibilizado pelo professor

Termo	Sinônimos	Valor
"Tomografia de crânio corte coronal"	"Tomo de crânio corte coronal", "TC de crânio corte coronal", "TC de seios da face", "tomografia de seios da face", "Tomografia de seios nasais", "Tomografia de face"	2,00
"Seio esfenoidal esquerdo"	"Seio esfenoidal do lado esquerdo", "Seio esfenoidal lateral esquerdo"	1,75
"aumento densidade partes moles"	"partes moles com maior densidade", "partes moles mais densas", "Aumento da atenuação", "densidade de partes moles", "partes moles densas", "focos hiperatenuantes"	1,50
"calcificações de permeio"	"permeio calcificado", "Calcificações esparsas", "áreas puntiformes", "áreas calcificadas", "calcificações", "calcificação de permeio"	1,50
"hipoaeração"	"hipo-aeração", "aerações reduzida", "pouco aerada"	1,50
"assimétrica"	"heterogênea", "expansão do seio"	1,00
"acometimento ósseo"	"destruição óssea", "remodelamento ósseo"	0,75

Tabela 5. Respostas de controle para a avaliação

Tipo de Resposta	Resposta
Gabarito completo	Tomografia de crânio corte coronal. Seio esfenoidal esquerdo, aumento de densidade nas partes moles, calcificações de permeio. Apresenta hipoaeração, revelando acometimento ósseo quase nulo.
Metade do gabarito	Tomografia de crânio corte coronal. Seio esfenoidal esquerdo, aumento de densidade nas partes moles.
Resposta incorreta	Não creio que essa informação seja pertinente para o diagnóstico.
Resposta vazia	

Tabela 6. Respostas de alunos para a avaliação

Identificação	Resposta
Aluno 1	Técnica: corte coronal de tomografia de seios da face, sem contraste iodado. Descrição: conteúdo no seio esfenoidal, com densidade de partes moles e focos de calcificação de permeio.
Aluno 2	Material com atenuação de partes moles com focos de calcificação no seu interior obliterando o seio esfenoidal esquerdo. Nota-se discreta esclerose óssea associada, sem lise.
Aluno 3	Conteúdo no seio esfenoidal esquerdo com áreas hiperatenuantes de permeio, sem aparente acometimento ósseo.
Aluno 4	Conteúdo hipodenso com focos de calcificações lineares no seio esfenoidal, sem efeito expansivo, com esclerose e espessamento de paredes osseas.
Aluno 5	Tomografia de seios nasais, reconstrução coronal, com lesão nodular no seio esfenoidal, hiperdensa, heterogênea, com focos de calcificação. Não se observa destruição ou reação óssea.
Aluno 6	Seio esfenoidal esquerdo repleto por conteúdo discretamente hiperdenso com focos de hiperdensidade de aspecto serpinginoso centrais.
Aluno 7	Seio esfenoidal esquerdo preenchido por material heterogêneo, destacando-se focos hiperatenuantes, no seu interior. Nota-se expansão deste seio, com deslocamento de suas paredes, sem a caracterização de destruição óssea.
Aluno 8	TC de seios da face sem contraste: Conteúdo heterogêneo no interior do seio esfenoidal, com atenuação de partes moles, observando-se calcificações associada e caráter expansivo com remodelamento ósseo.
Aluno 9	TC de face: lesão hipoatenuante em seio frontal com hiperdensidades centrais, sugestivas de calcificação.
Aluno 10	Conteúdo ocupando a porção esquerda do seio esfenoidal esquerdo, associada a focos hiperatenuantes produzindo pequeno efeito pequeno e discreto remodelamento ósseo adjacente, sem comprometimento evidente de estruturas adjacentes.

4. Resultados e discussões

Com base nas respostas apresentadas na seção 3, foram calculadas as notas e comparadas as correlações das mesmas com as notas atribuídas pelo professor, considerando a presença e a ausência de sinônimos.

4.1 Correção sem uso de sinônimos

Primeiramente as respostas fornecidas pelos alunos foram corrigidas somente com o pré-processamento de texto, sem usar sinônimos cadastrados pelo professor. A Tabela 7 mostra as notas obtidas com a avaliação de controle e a Tabela 8 apresenta as notas obtidas para as respostas reais dos alunos.

Conforme pode ser verificado na Tabela 7, a influência do pré-processamento é notória para a eficiência da correção, visto que o a correção alcançou as notas esperadas à medida que aumentou-se o nível de pré-processamento do sistema. Observa-se que houve apenas pequena diferença de nota em relação à questão “metade do gabarito”, devido à exatidão dos cálculos dos pesos pelo algoritmo. A tendência de correção manual é aproximar a nota final para valores inteiros, conforme pode ser observado na última coluna da Tabela 7. Um ponto importante a se ressaltar é que nesse momento, por serem os mesmos termos cadastrados pelo professor sendo usados como respostas de controle, foi obtida nota igual para os métodos com *stop words* e com *stemming*, ou seja, a forma de escrever empregada na resposta não afetou o algoritmo.

Tabela 7. Notas obtidas para a avaliação de controle sem uso de sinônimos.

Tipo de Resposta	Texto original	<i>stop words</i>	<i>Stemming</i>	Nota do Professor
Gabarito completo	5,24	10,00	10,00	10,00
Metade do gabarito	3,10	4,29	4,29	5,00
Resposta incorreta	0,00	0,00	0,00	0,00
Resposta vazia	0,00	0,00	0,00	0,00

Tabela 8. Notas obtidas para correção de respostas reais de alunos sem uso de sinônimos.

Identificação	Texto original	<i>stop words</i>	<i>Stemming</i>	Nota do Professor
Aluno 1	1,79	3,81	4,29	8,75
Aluno 2	1,19	2,14	2,86	7,50
Aluno 3	1,67	3,57	3,57	7,50
Aluno 4	0,00	1,67	2,38	7,50
Aluno 5	0,36	1,67	2,86	7,50
Aluno 6	1,43	1,43	1,43	7,50
Aluno 7	1,43	1,43	2,14	8,75
Aluno 8	1,07	3,10	3,57	10,00
Aluno 9	0,00	0,48	0,48	7,50
Aluno 10	1,19	2,14	2,14	8,75

A Tabela 8 indica que os valores das notas atribuídas para os alunos diferem substancialmente da nota do professor, mas há uma tendência de aproximação das notas à medida que se aumenta o pré-processamento.

Verifica-se na prática que o professor conhece termos correlatos que podem ser considerados durante a correção. Assim, esta abordagem de correção exata do texto afeta a acurácia da nota, visto que não considera este conhecimento do radiologista. Para mitigar tais problemas, como abreviações previstas, termos do dia-a-dia e as diversas formas que um termo pode ser escrito, a segunda abordagem considera o uso de sinônimos, conforme apresentado a seguir.

4.2 Correção com uso de sinônimos

O mesmo processo usado para a correção sem sinônimos foi repetido para avaliar as quatro respostas da avaliação de controle e as dez respostas reais de alunos. A Tabela 9 mostra as notas obtidas via sistema para as avaliações de controle e a Tabela 10 apresenta a correção para as respostas de alunos.

Tabela 9. Notas obtidas para a avaliação de controle com uso de sinônimos.

Tipo de Resposta	Texto original	<i>stop words</i>	<i>Stemming</i>	Nota do Professor
Gabarito completo	5,95	7,86	10,00	10,00
Metade do gabarito	3,10	5,00	5,00	5,00
Resposta incorreta	0,00	0,00	0,00	0,00
Resposta vazia	0,00	0,00	0,00	0,00

Como pode ser observado na Tabela 10, o uso de sinônimos favoreceu a aproximação das notas atribuídas pelo sistema às notas atribuídas pelo professor no que se refere à avaliação de controle, mesmo considerando os textos com menor pré-processamento (somente *stop words*).

Tabela 11. Notas obtidas para correção de respostas reais de alunos com uso de sinônimos.

Identificação	Texto original	<i>stop words</i>	<i>Stemming</i>	Nota do Professor
Aluno 1	3,10	5,95	7,14	8,75
Aluno 2	2,86	4,52	5,00	7,50
Aluno 3	3,81	5,00	5,48	7,50
Aluno 4	1,43	3,81	5,00	7,50
Aluno 5	2,38	6,67	7,14	7,50
Aluno 6	2,14	2,86	3,33	7,50
Aluno 7	2,14	5,71	6,67	8,75
Aluno 8	4,76	6,90	8,10	10,00
Aluno 9	1,19	2,14	2,62	7,50
Aluno 10	4,05	5,00	5,48	8,75

Os dados da Tabela 11 ratificam a importância do pré-processamento no processo de correção, pois o maior nível de processamento permite maior aproximação das notas corrigidas pelo sistema com as notas atribuídas pelo professor.

A diferença nos valores encontrados pela correção automática e pela nota dada pelo professor é decorrente principalmente devido à ausência de sinônimos não previstos pelo professor durante o cadastro da questão e o uso de abreviações. Observa-se, por exemplo, que o Aluno 9 incluiu o termo abreviado "Tomografia de seios" quando o correto seria "Tomografia de seios da face", resultando na diminuição da nota, uma vez que esses dois termos não foram fornecidos como sinônimos. Adicionalmente, foram verificados erros de digitação que dificultaram a recuperação dos sinônimos encontrados na base de dados. A correção de digitação e ortografia não fez parte do escopo do sistema nesta versão.

A fim de extrair uma medida de comparação, foi calculada a correlação entre os dois conjuntos de notas (sistema e professor), conforme mostra a Tabela 12.

Tabela 12. Correlações das notas obtidas na correção com uso de sinônimos.

Métodos Comparados	Correlação
Texto original X Professor	0,663
stop words X Professor	0,621
Stemming X Professor	0,682

A Tabela 12 mostra que apesar de as notas tenderem a crescer à medida que aumenta o pré-processamento, o método com uso de somente *stop words* não apresentou correlação muito maior que a correção com o texto original, como era esperado. Este fato ocorreu principalmente porque somente o uso *stop words* não foi capaz de conseguir identificar alguns termos, ação favorecida pelo processamento com *stemming*. O processo de *stemming* proporcionou um aumento da correlação, o que indica a consistência da correção, visto que a nota do sistema aumenta à medida que aumenta a nota atribuída pelo professor na correção manual. Acredita-se que o incremento de sinônimos não previstos no dicionário poderia aumentar ainda mais tal correlação.

Podemos exemplificar a importância da inserção de sinônimos pertinentes ao aplicar a correção com uso de um gabarito limitado (Tabela 13) e outro gabarito estendido (Tabela 4). Os resultados são comparados no gráfico da Figura 11, apresentando a correção com algoritmo com uso de *stemming*. Verifica-se que a correlação entre as duas notas (gabarito limitado e gabarito estendido) é positiva, mas a nota com o uso do dicionário estendido apresenta maior proximidade com a nota atribuída na correção manual pelo professor. A diferença entre as notas calculadas mostra que o conhecimento do assunto é necessário pelo agente especialista, no caso um radiologista, no momento do cadastro das questões e seus termos. Além disso, verifica-se a indicação do máximo possível de termos correlatos certos e errados contribui para a precisão da correção.

Tabela 13. Gabarito original.

Termo	Sinônimos	Valor
"Tomografia de crânio corte coronal"	"Tomo de crânio corte coronal", "TC de crânio corte coronal", "TC de seios da face", "tomografia de seios da face"	1,50
"Seio esfenoidal esquerdo"	"Seio esfenoidal do lado esquerdo", "Seio esfenoidal lateral esquerdo"	1,50
"aumento densidade partes moles"	"partes moles com maior densidade", "partes moles mais densas", "Aumento da atenuação"	1,00
"calcificações de permeio"	"permeio calcificado", "Calcificações esparsas", "áreas puntiformes", "áreas calcificadas", "calcificações"	1,50
"hipoaeração"	"hipo-aeração", "aerações reduzida", "pouco aerada"	1,50
"assimétrica"	"heterogênea"	1,00
"espessamento mucoso"	"espessamento da mucosa"	1,00
"acometimento ósseo"	"destruição óssea"	1,00

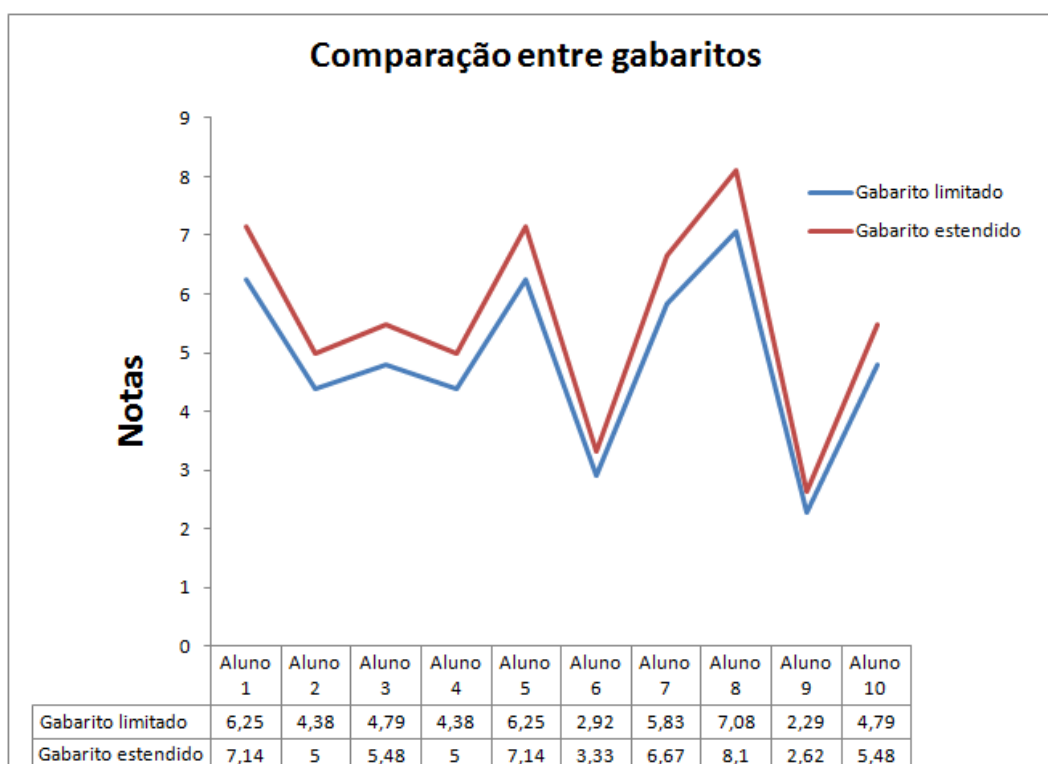


Figura 11. Comparação de resultados com uso de diferentes gabaritos.

O método apresentado neste trabalho tem por vantagens a facilidade, a flexibilidade e a agilidade na correção de avaliações dissertativas na área de Radiologia, o que pode contribuir para o aumento da qualidade na educação médica, visto que a aplicação de provas dissertativas pode ser mais adequada do que provas objetivas.

Os métodos aqui apresentados foram propostos para a área de Radiologia, como uma solução para uma área que atualmente se mostra carente em sistemas de informação voltados para o aprendizado, mas podem facilmente ser aplicados em outras áreas, considerando que faz uso de pré-processamento de texto. Da mesma forma, a abordagem pode ser adaptada para outras línguas, como a língua inglesa, por exemplo, sem grandes dificuldades, já que os processos de *stemming* de palavras e de inclusão de sinônimos podem ser facilmente aplicados a outras línguas.

O resultado apresentado se mostra satisfatório, considerando a simplicidade e a rapidez no processo de correção. Verifica-se adicionalmente que a inclusão de sinônimos que representem de forma mais adequada o conhecimento do professor pode melhorar a eficiência da correção. Essas inclusões podem até mesmo ser conduzidas interativamente, pois à medida que o professor avalia o resultado do sistema como não adequado, o mesmo pode aperfeiçoar o dicionário com a inclusão de termos certos e errados que deverão ser observados no algoritmo de correção. Para isso, ele pode incluir novos sinônimos e reprocessar a correção de forma bastante prática usando o sistema *RadioAssess*.

5. Conclusão e trabalhos futuros

O sistema *RadioAssess* e os algoritmos de correção se mostraram plenamente consistentes, tendo sido comprovados pelo estudo de caso de controle e pelo estudo de caso com respostas reais. Verifica-se que há uma dificuldade na representação do conhecimento do professor, o que exerce influência direta na correlação entre as notas atribuídas pelo professor e as notas atribuídas de forma automatizada.

As técnicas de pré-processamento de *stop words* e *stemming* implementadas mostraram-se essenciais para que a correção seja eficiente. A abordagem de sinônimos apresentada neste trabalho mostrou-se eficiente e flexível, visto que o professor pode incrementar o seu uso à medida que verifica distanciamento entre as notas atribuídas pelo sistema e as notas esperadas por ele. A modelagem de dados apresentada permite definição irrestrita de sinônimos e abreviações, constituindo uma grande vantagem do sistema *RadioAssess*.

Adicionalmente, o sistema apresentado possibilita a elaboração de provas diferentes com as mesmas questões, apoiando a correção do professor. Possibilita, ainda, a correção personalizada para cada professor, isto é, cada professor pode definir um conjunto de sinônimos diferentes para cada termo de cada questão da avaliação. Esta característica garante flexibilidade ao processo de correção, possibilitando a representação do conhecimento de forma mais adequada para cada docente usuário do sistema.

Outras técnicas podem ser aliadas à presente proposta para que melhores resultados sejam obtidos, como por exemplo uso de interpretação de linguagem natural associada ao tratamento com sinônimos.

Por fim, ressalta-se que além da facilidade de uso, flexibilidade e agilidade da correção oferecidas pelo sistema *RadioAssess*, este pode constituir-se em uma ferramenta de autoavaliação do Professor, visto que as notas atribuídas pelo sistema podem constituir um subsídio de retroalimentação a fim de que o docente confira a correção feita manualmente.

References

- The Apache Software Foundation, "Welcome to The Apache Software Foundation!", Available at <http://www.apache.org/>, October.
- Arnold L. B., Usery J. B., Finch C. K., Wallace J. L., Deaton P. R. and Self T. H. (2009) "Inadequate Documentation of Asthma Management in Hospitalized Adult Patients", In: Southern Medical Association, pages 510-514 102(5)
- Buriasco, R. L. C. (2000) "Algumas considerações sobre avaliação educacional", In: Avaliação Educacional, pages 155-178 (22)
- Dikli, S (2006) "An Overview of Automated Scoring of Essays", In: The Journal of Technology, Learning, and Assessment, vol. 5, n. 1.
- Evans, C. "The Official YAML Web Site", Available at <http://www.yaml.org/>, August.
- Evans, C., Ben-Kiki, O., döt-Net, I. "YAML Ain't Markup Language (YAML™) Version 1.2", Available at <http://www.yaml.org/spec/1.2/spec.html>, August.
- Grossman, E., Cardoso, M. H. C. A. (2006) "As narrativas em medicina contribuições a pratica clínica e ao ensino medico", In: Revista Brasileira de Educação Médica, pages 6-14 30(1)
- Hauben, M. and Reich, L. (2005) "Communication of findings in pharmacovigilance: use of the term "signal" and the need of precision in its use", In: European Journal of Clinical Pharmacology, pages 479-480 61(5)
- Kaur, N. and Jyoti, K. (2012) "Automated Assessment of Short One-Line Free-Text Responses In Computer Science", In: Special Issue of International Journal of Computer Science & Informatics (IJCSI), ISSN (PRINT) : 2231–5292, Vol. II, Issue 2, pages 105-109 2(1)
- Mishra, A. K., Bhattarai, S., Bhurtel, P, Bista, N. R., Shrestha, P., Thakali, K., Banthia, P. and Pathak, S. R. (2009) "Need for Improvement of Medical Records", In: J Nepal Med Assoc 2009, pages 10-3-6 48(174)
- Oracle Corporation "MySQL 5.6 Reference Manual", Available at <http://dev.mysql.com/doc/refman/5.6/en/>, August.
- Orengo, V.M. and Huyck, C. (2001) "A Stemming Algorithm for the Portuguese Language", In: Proceedings of the Eighth International Symposium on String Processing and Information Retrieval, pages 186-193
- PHP Group. "PHP Hypertext Preprocessor", Available at <http://php.net/>, August.

- Potencier, F. "What is Symfony", Available at <http://symfony.com/what-is-symfony>, August.
- Salton, G. D. and Carniel, C. A. (2013) "Regras sintáticas livres de contexto na correção de Unidades de Leitura", In: Nuevas Ideas en Informática Educativa TISE 2013, pages 217-222
- Santos, T. L. T., Silva, A. S., Favero, E. L. and Lino, A. D. P. (2007) "Avaliação automática de questões conceituais discursivas", In: IX Argentine Symposium on Artificial Intelligence, ASAI, 2007, Mar Del Plata. 36ª Jornadas Argentinas de Informática, JAIIO. Buenos Aires: Sociedad Argentina de Informática - SADIO, 2007, pages 128-138
- Valenti, S. Neri, F. Cucchiarelli R (2003) "An overview of current research on automated essay grading", In: Journal of Information Technology Education, vol. 2, pages 2003-2015.
- World Wide Web Consortium "HTML Current Status - W3C" Available at http://www.w3.org/standards/techs/html#w3c_all, August.
- World Wide Web Consortium "CSS Current Status - W3C" Available at http://www.w3.org/standards/techs/css#w3c_all, August.
- WebPlatform.org "javascript · javascript · WPD · WebPlatform.org", Available at <http://docs.webplatform.org/wiki/javascript>, August.