TÍTULO

Avaliação dos sistemas de correção automática de redação disponíveis para o português do Brasil

TEMA E DELIMITAÇÃO DO ESTUDO

O presente projeto se inscreve no campo da "Correção Automática de Redações" (em inglês: *Automated Essay Escoring*, ou AES) e da "Avaliação Automática de Redações" (*Automated Essay Evaluation*, ou AEE). Trata-se de subdomínio do processamento automático das línguas naturais (PLN), que explora o uso de tecnologias de inteligência artificial para a automatização do processo de identificação de problemas e atribuição de notas aos textos normalmente desenvolvidos em ambiente escolar. O projeto se situa, portanto, na região de encontro das ciências da linguagem e das ciências da computação e tem caráter eminentemente interdisciplinar, voltando-se, especificamente, para o recenseamento dos sistemas e tecnologias empregados para a correção automática de redações para o português brasileiro, bem como para a avaliação de desempenho das ferramentas disponíveis.

PROBLEMA DE TRABALHO

O projeto se organiza em torno de um problema de natureza prática: Quão eficientes são os sistemas de correção automática de redações disponíveis para o português brasileiro? Esse problema de trabalho se subdivide em três subproblemas conjugados, que serão endereçados sequencialmente durante o desenvolvimento da pesquisa:

a) Quais são os sistemas de correção automática de redações disponíveis para o português brasileiro?
b) Qual é a medida de "eficiência" de um sistema de correção automática de redações, consideradas, particularmente, as expectativas da comunidade de usuários (professores, estudantes e sistemas de ensino)?
c) Como avaliar a "eficiência" de um sistema de correção automática de redações, dadas as métricas de avaliação resultantes do subproblema anterior?

HIPÓTESE

O projeto parte da hipótese de que o desempenho dos sistemas de correção automática de redações disponíveis para o português brasileiro está ainda muito aquém da expectativa dos usuários. Os sistemas existentes seriam ainda caracterizados por inúmeras limitações, entre as quais:

a) a não detecção de muitos dos desvios do texto normalmente identificáveis por um corretor humano (falsos negativos);
b) a sobrecorreção do texto, com a identificação de desvios inexistentes (as chamadas "alucinações" de sistemas de inteligência artificial, ou falsos positivos);
c) a atribuição de notas discrepantes das atribuídas pelo avaliador humano; e
d) o fornecimento de feedbacks insuficientes ou enganadores para o usuário.

Essas hipóteses serão verificadas, durante o curso do projeto, por meio de avaliação quantitativa e qualitativa do desempenho dos sistemas para um *corpus* de testes constituído

por redações previamente corrigidas por avaliadores humanos, que será tomado como parâmetro para a comparação.

OBJETIVOS GERAL E ESPECÍFICOS

O objetivo geral do projeto é identificar o grau de precisão e confiabilidade dos sistemas de correção automática de redações disponíveis para o português brasileiro. A consecução desse objetivo geral depende da consecução de quatro objetivos específicos, abaixo assinalados:

a) Identificar os sistemas de correção automática de redações disponíveis para o português brasileiro;
b) Construir métricas de avaliação para sistemas de correção automática que possam ser utilizadas para testar os sistemas identificados;
c) Testar os sistemas identificados a partir das métricas elaboradas, tendo como parâmetro de referência o desempenho dos corretores humanos;
d) Analisar os resultados, de forma a indicar as limitações e possíveis vias de resolução dos problemas identificados.

JUSTIFICATIVA

A relevância da pesquisa em correção automática de redações pode ser sustentada em duas perspectivas principais: a social e a teórica.

A Base Nacional Curricular Comum (Resoluções CNE/CP n° 2/2017 e n°4/2018) estabelece a Produção de Textos como um dos quatro eixos em que se organiza o componente de Língua Portuguesa na Educação Básica. A orientação corrobora a perspectiva das diretrizes e dos parâmetros curriculares nacionais para a Educação Básica (Resolução CNE/CP n° 4/2010), que já enfatizavam o texto como a principal unidade de ensino de Língua Portuguesa, e que recomendavam a prática frequente e sistemática de produção de textos como estratégia indispensável ao desenvolvimento das habilidades de escrita essenciais à formação do estudante.

No entanto, a correção de redações é uma atividade escolar extremamente laboriosa. Segundo Bittencourt (2020, p. 19), um professor leva, em média, 12 minutos para atribuir nota a uma redação do Exame Nacional do Ensino Médio (ENEM). Se somarmos, a esse número, também o tempo necessário para a identificação dos problemas e o fornecimento de feedback individualizado para o estudante, percebe-se que a correção de redações é uma prática não escalável e pouco adaptada à realidade dos sistemas de ensino que contam com salas de aulas numerosas e dedicação apenas parcial dos professores, que normalmente se desdobram em diferentes turmas e unidades curriculares. Por essa razão, os alunos têm tido poucas oportunidades de produção de textos na educação básica, principalmente na rede pública de ensino, situação que termina por dificultar o desenvolvimento das habilidades de escrita esperadas da formação escolar.

O emprego de sistemas automáticos de correção de redações como ferramentas auxiliares de ensino – asseguradas sua eficiência, precisão e credibilidade – poderia contribuir positivamente para a transformação desse cenário, estimulando a produção mais intensiva e mais frequente de textos sem sobrecarga docente e com resultados mais rápidos.

Adicionalmente, a automação do processo de correção poderia contribuir para a redução das despesas e do tempo envolvido em avaliações gerais, caso do ENEM, cujo orçamento, em 2023, teria custado R$ 346,2 milhões aos cofres públicos (dos quais, estima-se, um terço tenha sido empregado na correção das redações). Tome-se, como exemplo, o

TOEFL (*Test of English as a Foreign Language*), exame requerido para o ingresso de estrangeiros em universidades americanas, cuja nota final é a média aritmética simples das correções humana e automática (cabendo segunda correção humana nos casos de divergência expressiva entre as duas avaliações)[1].

A par dos evidentes benefícios sociais, ressalte-se a relevância teórica da investigação. A correção de redações é atividade complexa, que envolve vários níveis de investigação linguística: desde o reconhecimento da caligrafia empregada até a avaliação global dos textos, passando por inúmeros estágios intermediários, como o processamento ortográfico e morfossintático (para a identificação de desvios à norma-padrão da língua portuguesa), o processamento semântico-pragmático (para a identificação de problemas nos processos de constituição de sentidos), o processamento textual (para a avaliação dos mecanismos de organização do texto, principalmente os de coesão e coerência) e o processamento discursivo (para avaliação, entre outras, da adequação temática, da tipologia textual, da qualidade do repertório e das relações intertextuais).

A correção de redações é, portanto, um domínio privilegiado de pesquisa teórica, que permite a consideração simultânea e integrada de fenômenos linguísticos de natureza muito variada, e cuja exploração pode fazer avançar consideravelmente o estado da arte do conhecimento humano sobre a linguagem. A correção automática, por sua vez, na medida em que busca emular computacionalmente as competências linguísticas envolvidas na atividade de avaliação de textos, é uma instância extraordinária de validação de hipóteses, e pode promover o desenvolvimento de tecnologias de processamento linguístico que já vêm sendo empregadas, com resultados promissores, em vários outros domínios, como o reconhecimento de voz, a síntese de fala e a tradução automática, para citar três exemplos de sistemas hoje disponíveis ao público.

DISCUSSÃO TEÓRICA E METODOLÓGICA

O campo da correção automática de redações vem sendo explorado sistematicamente pelo menos desde a década de 1960[2], mas as tentativas de desenvolvimento de sistemas para o português brasileiro são relativamente recentes[3]. Como as práticas de correção de textos são culturalmente orientadas e dependentes de língua, as possibilidades de transferência tecnológica do inglês para o português são muito limitadas, e envolvem o desenvolvimento de sistemas próprios, que deem conta, não apenas da variedade de critérios utilizados pelas instituições brasileiras (ENEM, Fuvest, Unesp, Unicamp, UnB, etc.), mas também dos fenômenos característicos da língua portuguesa.

Essa especificidade não impediu, porém, o desenvolvimento de sistemas próprios para correção de redações em português. Em levantamento preliminar, reconhecem-se já pelo menos quatro sistemas disponíveis no mercado brasileiro:

- Coredação (https://coredacao.com/)
- CIRA (http://www.ciraredacoes.com.br/)
- CRIA (https://cria.net.br/)
- Glau (https://www.glau.com.vc/)

[1] Bridgeman (2013)

[2] Shermis, Burstein (2013)

[3] Rassi, Lopes (2023); Lima (2023)

A essas quatro aplicações soma-se a plataforma Redação SP, lançada pela Secretaria de Educação do Estado de São Paulo em agosto de 2023, mas de uso interno e exclusivo aos professores da rede pública paulista.

Todas essas aplicações afirmam empregar técnicas de inteligência artificial, e o material publicitário disponível em cada caso leva a crer que esses sistemas já teriam alcançado precisão comparável à de um corretor humano. Com efeito, os exemplos utilizados para promover cada uma dessas ferramentas são surpreendentes e provocam a impressão de que a correção automática de redações é um problema resolvido.

No entanto, e como se trata de sistemas computacionais protegidos por segredo comercial, não é fornecida a documentação relativa aos algoritmos empregados, e tampouco são disponibilizados resultados de testes comparativos isentos que possam comprovar o real alcance e a verdadeira precisão dessas aplicações. Um dos objetivos deste projeto é exatamente prover essa lacuna: construir um *benchmarking* que possa ser utilizado para aferir o grau de credibilidade e confiabilidade do desempenho dos sistemas disponíveis de correção automática de redações, e que possa ser empregado como parâmetro de avaliação de desenvolvimentos futuros.

A suspeita é de que o material utilizado para a divulgação dos sistemas – e a venda de planos de assinatura correspondentes – seja excessivamente otimista.

Os vários sistemas de correção automática de redações desenvolvidos em âmbito acadêmico, e que são, por definição, transparentes quanto à codificação e sujeitos à avaliação dos pares, apresentam resultados bem mais modestos e apontam limitações que estariam escamoteadas nos apelos publicitários dos sistemas comerciais. Percebe-se, por exemplo, que os sistemas acadêmicos vêm se desenvolvendo sobre competências isoladas de avaliação, cada uma envolvendo problemas próprios, com técnicas particulares, que longe estão de uma perspectiva de solução.

Alguns sistemas empregam, por exemplo, ferramentas estatísticas – como Coh-Metrix e Linguistic Inquiry Word Count (LIWC) – adaptadas para o português para a extração de índices de legibilidade e coesão, baseados principalmente na quantidade e na qualidade do vocabulário empregado nas redações[4]. Outros trabalhos utilizam métricas calculadas a partir da similaridade de cosseno para avaliar a adequação da redação ao tema solicitado[5]. Algumas iniciativas operam a partir da mineração de argumentos para avaliar a construção da argumentação[6]. Em todos esses casos, o processo de avaliação de redações está ainda encapsulado em torno de critérios especializados, não sendo cabível ainda a atribuição de uma nota global à redação.

É pouco crível que os sistemas comerciais possam alcançar resultados tão mais avançados do que os descritos pela pesquisa científica, principalmente se considerado o fato de que essa pesquisa se vale, para a extração de atributos, de modelos de classificação e regressão treinados a partir de algoritmos dominantes na área de inteligência artificial: regressão linear, *Support Vector Machines*, *Gradient Boosting*, *Transformers*, etc[7].

Também é pouco razoável supor que os mesmos sistemas comerciais possam ter se valido de quantidades de dados muito superiores às empregadas na pesquisa científica, que oscila entre conjuntos de pouco mais de 56.000 redações, com diversidade limitada de temas,

---

[4] Ferreira et al. (2021)

[5] Amorim, Veloso (2017)

[6] Sousa et al (2021)

[7] Rassi, Lopes (2023)

o que representa um obstáculo para o treinamento de modelos utilizando técnicas de *deep learning*.

Paira, portanto, sobre os sistemas comerciais, uma desconfiança que se pretende, neste projeto, colocar à prova. Para que se possa superar a opacidade da documentação disponível e validar (ou não) os resultados por eles apresentados, esta proposta percorre um caminho metodológico organizado em cinco momentos:

1) Recenseamento dos sistemas disponíveis de correção automática de redações para o português brasileiro. Embora já tenham sido identificados quatro sistemas abertos ao público, o primeiro movimento deste trabalho será empreender uma busca mais informada para averiguar se não haveria outras aplicações com os mesmos objetivos que tenham escapado ao levantamento preliminar. Da mesma forma, pretende-se fazer uma busca ativa por documentação dos sistemas disponíveis, possivelmente com entrevistas (a distância) com os responsáveis pelo desenvolvimento, para que se possa inferir sobre características de implementação que possam ser utilizadas, mais adiante, para análise do desempenho observado. No entanto, dada a natureza comercial das ferramentas, não se espera, pelo menos em princípio, que dessa busca ativa venham a resultar informações muito relevantes.

2) Constituição do *corpus* de testes. Para que se possa avaliar comparativamente o desempenho dos sistemas será compilado um conjunto de redações, o mais próximo possível de situações espontâneas de produção (contexto escolar), sobre temas variados, que serão posteriormente submetidas à correção humana, por corretores treinados e capacitados, cujas observações (erros observados, notas atribuídas, comentários emitidos) serão utilizadas como parâmetro de avaliação dos sistemas sob escrutínio. A princípio, estima-se que, nessa etapa, sejam coligidas 10 redações, de 5 temas diferentes, a serem submetidas à correção cega por pelo menos três corretores humanos cada uma. Idealmente, as correções observarão os critérios adotados pelo ENEM, além da identificação dos erros e do *feedback* para o aluno, dado que todos os sistemas listados operam, pelo menos, com esses parâmetros. As correções humanas serão comparadas e serão identificadas as diferenças de desempenho entre os corretores: discrepâncias entre notas para cada uma das cinco competências do ENEM; erros identificados por apenas um subconjunto dos corretores; natureza dos comentários em cada caso.

3) Avaliação dos sistemas. No terceiro momento, o *corpus* de testes será submetido à correção automática dos sistemas listados. Os resultados serão comparados e serão assinaladas as diferenças de desempenho entre os sistemas automáticos, e entre os sistemas automáticos e os corretores humanos.

4) Análise dos resultados. Os resultados das fases anteriores serão tabulados e analisados, de forma a inspirar a constituição de métricas de avaliação que possam ser utilizadas para quantificar e qualificar as diferenças de desempenho observadas.

5) Interpretação dos resultados. A partir da análise dos resultados dos sistemas automáticos se buscará operar uma engenharia reversa que permita identificar a natureza das técnicas de implementação empregadas, sempre que possível.

CRONOGRAMA

Os momentos indicados na seção anterior serão distribuídos ao longo de 12 meses duração do projeto, segundo o cronograma provável abaixo:

Mês	Atividade	Resultado Esperado
1	Recenseamento dos sistemas de correção automática de redações disponíveis para o português brasileiro, bem como o levantamento da documentação técnica correspondente.	Lista de sistemas a serem testados
2	Compilação do corpus de testes (conjunto de redações produzidas por estudantes da educação básica em situações convencionais de produção de texto)	Corpus de 10 redações, de 5 temas diferentes
3	Correção humana do corpus de testes por pelo menos três corretores capacitados (que tenham sido aprovados pelo curso de preparação de corretores e que tenham experiência prévia de correção de redações do ENEM)	Corpus corrigido por humanos (3 x 10 redações)
4	Correção automática dos textos (submissão do corpus de testes aos sistemas de correção a serem avaliados)	Corpus corrigido por sistemas de IA
5	Análise comparativa das correções da competência 1 do ENEM (Língua Portuguesa)	Resultado da análise comparativa
6	Análise comparativa das correções da competência 2 do ENEM (Abordagem temática e adequação ao tipo textual)	Resultado da análise comparativa
7	Análise comparativa das correções da competências 3 do ENEM (Progressão textual e defesa do ponto de vista)	Resultado da análise comparativa
8	Análise comparativa das correções da competeência 4 do ENEM (Coesão e articulação)	Resultado da análise comparativa
9	Análise comparativa das correções da competência 5 do ENEM (Proposta de intervenção)	Resultado da análise comparativa
10	Consolidação e tabulação dos dados, com a definição de métricas de comparação	Quadro comparativo e métrica de avaliação
11	Interpretação dos resultados e tentativa de inferência dos modelos empregados pelos sistemas avaliados	Conclusões da avaliação
12	Redação do relatório final da pesquisa	Relatório final

REFERÊNCIAS

AMORIM, E.; VELOSO, A. A Multi-aspect Analysis of Automatic Essay Scoring for Brazilian Portuguese. Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics. Anais...Valencia, Spain: Association for Computational Linguistics, abr. 2017.

BITTENCOURT JR., J. A. S. Avaliação automática de redação em língua portuguesa empregando redes neurais profundas. Universidade Federal de Goiás, 2020.

BRIDGEMAN, B. Handbook of automated essay evaluation: Current applications and new directions. Em: SHERMIS, M. D.; BURSTEIN, J. (Eds.). [s.l.] Routledge/Taylor & Francis Group, 2013. p. 221–232.

DA SILVA JR., J. A. Um avaliador automático de redações. Universidade Federal do Espírito Santo, 2021.

FERREIRA MELLO, R. et al. Towards automated content analysis of rhetorical structure of written essays using sequential content-independent features in ortuguese. (A. F. Wise, R. Martinez-Maldonado, I. Hilliger, Eds.) LAK22 Conference Proceedings. Anais...United States of America: Association for Computing Machinery (ACM), 2022.

FERREIRA, R. et al. Towards Automatic Content Analysis of Rhetorical Structure in Brazilian College Entrance Essays. Em: [s.l: s.n.]. p. 162–167.

FONSECA, E. R. et al. Automatically Grading Brazilian Student Essays. (A. Villavicencio et al., Eds.) Computational Processing of the Portuguese Language. Anais...Springer International Publishing, 2018.

HAENDCHEN FILHO, A. et al. An approach to evaluate adherence to the theme and the argumentative structure of essays. International Conference on KnowledgeBased Intelligent Information & Engineering Systems. Anais...2018.

HAENDCHEN FILHO, A. et al. Imbalanced Learning Techniques for Improving the Performance of Statistical Models in Automated Essay Scoring. Procedia Computer Science, v. 159, p. 764–773, jan. 2019.

LIMA, T. B. DE et al. Avaliação Automática de Redação: Uma revisáo sistemática. Revista Brasileira de Informática na Educação, v. 31, p. 205--221, maio 2023.

MARINHO, J. et al. Automated Essay Scoring: An approach based on ENEM competencies. Anais do XIX Encontro Nacional de Inteligência Artificial e Computacional. Anais...SBC, 2022.

MARINHO, J.; ANCHIÊTA, R.; MOURA, R. Essay-BR: a Brazilian Corpus to Automatic Essay Scoring Task. Journal of Information and Data Management, v. 13, n. 1, p. 65–76, 2022.

SHERMIS, M. D.; BURSTEIN, J. Handbook of Automated Essay Evaluation: Current Applications and New Directions. [s.l.] Routledge/Taylor & Francis Group, 2013.

SOUSA, A. et al. Cross-Lingual Annotation Projection for Argument Mining in portuguese. (G. Marreiros et al., Eds.)Progress in Artificial Intelligence. Anais...Springer International Publishing, 2021.