Análise automatizada de coesão em redações do ensino fundamental por meio de técnicas de processamento de linguagem natural

Aluno: Bruno Alexandre Barreiros Rosa[1]
Orientador: Rafael Ferreira Mello[1]

[1]Centro de Estudos e Sistemas Avançados do Recife (CESAR SCHOOL)
Recife – PE – Brasil

babr@cesar.school, rflm@cesar.school

Resumo. *A correção de redação é um trabalho manual recorrente do professor para avaliar o domínio da escrita do aluno na educação básica. A coesão é um aspecto fundamental do texto, visto que auxilia na relação de sentido. Tratar de forma automática a coesão em redações da língua portuguesa é um desafio na área de Processamento de Linguagem Natural (PLN). O objetivo da pesquisa é propor um algoritmo para analisar elementos de coesão em redações dos anos finais do ensino fundamental brasileiro por meio de técnicas de PLN. A pesquisa experimental será realizada em uma base de dados real com cerca de 2.000 redações de escolas participantes das avaliações formativas do programa Brasil na Escola do Ministério da Educação (MEC).*

1. Introdução

A escrita é uma habilidade fundamental para viver em sociedade, conforme reconhecido pela UNESCO [UNESCO 2011]. É, também, uma habilidade fundamental para o indivíduo obter sucesso no mundo acadêmico, corporativo ou, até mesmo, na vivência social [Graham 2019]. A redação é considerada um mecanismo útil para avaliar os resultados da aprendizagem em escrita, orientar o processo de aprendizagem e medir o progresso de estudantes [Zupanc and Bosnić 2017]. Entretanto, a correção manual de redação é uma tarefa que requer muito tempo e a disponibilidade da maioria dos educadores é limitada [Correnti et al. 2020]. Além disso, durante a correção de grande volume de redações, o educador dificilmente consegue manter o mesmo nível em todos os textos [Graham 2019]. Assim, o desenvolvimento de ferramentas para auxiliar a correção de atividades de escrita de estudantes ainda é um desafio [Ifenthaler 2022].

No Brasil, apesar da relevância das habilidades de escrita, muitas escolas da educação básica não têm conseguido estimular adequadamente essa capacidade, como visto nos resultados de avaliações em larga escala realizadas nos últimos anos [Passero et al. 2019]. Nessa perspectiva, o Ministério da Educação (MEC) publicou o decreto nº 11.079[1], que institui a Política Nacional para Recuperação das Aprendizagens na Educação Básica, visando implementar estratégias, programas e ações para a recuperação das aprendizagens e o enfrentamento da evasão e do abandono escolar. Com isso, um dos eixos do decreto propõe o fortalecimento e expansão das práticas e das abordagens educacionais existentes por meio do uso de novas tecnologias e de recursos digitais.

[1]Decreto nº 11.079, de 23 de maio de 2022: https://www.in.gov.br/web/dou/-/decreto-n-11.079-de-23-de-maio-de-2022-402040949

DOI: 10.5753/cbie_estendido.2022.226737

A estratégia do Programa Brasil na Escola, instituído pelo MEC, é possibilitar o diagnóstico das aprendizagens e o mapeamento dos estudantes com mais dificuldades, permitindo a personalização no acompanhamento e intervenções pedagógicas, diminuindo as desigualdades na sala de aula e nas unidades escolares. O programa permite, em um dos seus eixos, que professores apliquem avaliações de redações formativas para os alunos dos anos finais do Ensino Fundamental. Após a aplicação, o professor poderá submeter a folha de resposta da redação no aplicativo Plataforma de Digitalização[2], a fim de que as correções das redações possam ser realizadas considerando quatro aspectos: (1) Registro, (2) Coerência Temática, (3) Tipologia Textual e (4) Coesão.

Nesse sentido, muitos pesquisadores estão trabalhando na análise automatizada de redação utilizando técnicas de Processamento de Linguagem Natural (PLN) para automatizar o processo [Ferreira Mello et al. 2019]. As soluções desenvolvidas objetivaram basicamente atender às questões de tempo, custo, confiabilidade e aplicabilidade com relação à avaliação da escrita [Ifenthaler 2022]. Embora as pesquisas se concentraram em sua maioria na língua inglesa [Ifenthaler 2022], outros idiomas iniciaram seus desenvolvimentos, como por exemplo, o Árabe [Azmi et al. 2019], o Alemão [Pirnay-Dummer and Ifenthaler 2011], o Hebraico [Cohen et al. 2003] e o Suéco [Östling et al. 2013].

Avaliar uma redação considerando todos os parâmetros, como a relevância do conteúdo para o enunciado, o desenvolvimento de ideias, a coesão e a coerência, é um grande desafio até o momento [Ramesh and Sanampudi 2022, Ifenthaler 2022]. A principal fraqueza é o foco predominante no vocabulário, na sintaxe e na consideração limitada da semântica do texto [Ramesh and Sanampudi 2022]. [Halliday and Hasan 1995] dizem que um texto é mais bem considerado como uma unidade semântica: uma unidade não de forma, mas de relações de sentido, que se realizam por intermédio de construções linguísticas. Nessa perspectiva, a coesão é a propriedade pela qual se cria e se sinaliza toda espécie de ligação, de laço, que dá ao texto unidade de sentido ou unidade semântica [Antunes 2005].

Em inglês, existem abordagens promissoras para coesão que foram propostas [Tian et al. 2021]. Na língua portuguesa, ainda é um desafio tratar elementos coesivos em redações de forma automática [Grama 2022, Lima et al. 2018]. O projeto Brasil na Escola, nos anos finais do ensino fundamental, trabalha a coesão com a finalidade de avaliar o domínio dos mecanismos linguísticos necessários para a construção da narrativa, ou seja, a conexão, a ligação e a harmonia entre os elementos linguísticos de um texto. Nessa etapa de escolaridade, não se espera o domínio de estruturas complexas previstas para etapas subsequentes, mas o conhecimento do estudante sobre a necessidade de interligar as partes do seu texto de forma linear, do ponto de vista da referenciação e sequenciação.

A partir do exposto acima, busca-se responder à seguinte pergunta de pesquisa: *Como automatizar a análise de coesão em redações da língua portuguesa dos anos finais do ensino fundamental para apoiar a correção do professor por meio de técnicas de processamento de linguagem natural?*

[2]Acesso em 20/07/2022: `https://www.gov.br/mec/pt-br/acesso-a-informacao/institucional/secretarias/secretaria-de-educacao-basica/programas-e-acoes/plataforma-de-avaliacoes-diagnosticas-e-formativas`

Visando responder a essa pergunta de pesquisa, foram estabelecidos o objetivo geral e os específicos, descritos a seguir:

Objetivo Geral:

- Propor um algoritmo para analisar elementos de coesão em redações da língua portuguesa dos anos finais do ensino fundamental para apoiar a correção do professor por meio de técnicas de Processamento de Linguagem Natural (PLN).

Objetivos Específicos:

- Realizar uma revisão sistemática da literatura sobre a aplicação de técnicas de PLN em atividades de escrita na educação básica brasileira;
- Propor e desenvolver um algoritmo para avaliar elementos coesivos em produções textuais utilizando técnicas de processamento de linguagem natural;
- Analisar o modelo utilizando um banco de dados real de redações da língua portuguesa.

2. Metodologia

Para atingir os objetivos propostos, a pesquisa experimental será desenvolvida utilizando métodos quantitativos por meio das seguintes etapas: (1) Revisão Sistemática da Literatura; (2) Definição dos atributos de coesão; (3) Criação do banco de dados de redações; (4) Processamento do algoritmo; e (5) Avaliação dos resultados.

2.1. Revisão Sistemática da Literatura

A Revisão Sistemática da Literatura (RSL) está em fase final e tem o propósito de demonstrar o estado da arte da automatização de correção de redações e questões abertas da língua portuguesa que utilizaram técnicas de Processamento de Linguagem Natural (PLN) no contexto da educação básica.

Nessa perspectiva, foram definidas as perguntas de pesquisa, a *string* de busca e os critérios de inclusão e exclusão. As buscas foram realizadas em bases de dados renomadas, a fim de encontrar o estado da arte entre o período de 2012 a 2022, que resultaram em 1.056 artigos. A ferramenta Rayyan foi utilizada para rotular os critérios de inclusão e exclusão, sendo selecionados, ao final, 8 artigos. Contudo, no intuito de ampliar o escopo de estudos primários, a técnica *Snowballing* foi aplicada, resultando em 19 artigos. A conclusão da RSL resultará em um artigo que será submetido a um periódico para possível publicação.

2.2. Definição dos Atributos de Coesão

A coesão textual refere-se ao uso de vocabulário e estruturas gramaticais para conectar as ideias contidas em um texto [Koch 1989]. Os mecanismos de coesão podem ser divididos em dois grupos: referencial e sequencial [Koch 1989]. O referencial engloba o uso de elementos que recupera e introduz um assunto ou algo que está presente no texto (referência endofórica) ou fora do texto (referência exofórica). Já o sequencial engloba o uso de elementos que estabelecem entre os segmentos do texto diversos tipos de relações semânticas, demonstrando sequencialidade às ideias apresentadas.

Para analisar a coesão textual, será considerada a matriz de correção estabelecida na avaliação formativa de escrita do programa Brasil na Escola. Na matriz, a avaliação dos aspectos coesivos de uma redação varia do Nível I ao V, descritos a seguir:

- **Nível I:** Palavras e períodos justapostos e desconexos ao longo do texto, ou seja, ausência de articulação, porém há uma coesão marcada pela relação lógica entre palavras e/ou enunciados ou repertório coesivo escasso e com desvios recorrentes;
- **Nível II:** Repertório coesivo escasso, mas com desvios pontuais;
- **Nível III:** Repertório coesivo pouco diversificado com recorrência de 1 (um) tipo de desvio;
- **Nível IV:** Repertório coesivo diversificado com desvios pontuais que ainda afetam a inteligibilidade de parte do texto;
- **Nível V:** Repertório coesivo diversificado com a possibilidade de raras inadequações que não interferem na inteligibilidade textual.

2.3. Criação do banco de dados de redações

O banco de dados de redações será composto por aproximadamente 2.000 textos narrativos, que foram realizados por alunos dos anos finais do ensino fundamental de escolas públicas brasileiras. Essas redações serão avaliadas previamente por dois professores humanos levando em consideração os critérios e níveis de correção dos elementos de coesão estabelecidos nas avaliações formativas do programa Brasil na Escola. Cabe ressaltar que cada professor não conhece a pontuação do outro e, caso haja divergência entre as duas pontuações, um terceiro professor avaliador atribuirá sua classificação. Assim, a base dados para a pesquisa será constituída.

2.4. Processamento do algoritmo

A proposta desta pesquisa volta-se para a aplicação de técnicas de PLN, utilizando-se algoritmos para mensurar os elementos de coesão em redações do ensino fundamental brasileiro. No pré-processamento, busca-se uma representação com apenas informações relevantes para o processo de avaliação. Nessa fase, utilizam-se técnicas de PLN para normalização das redações. [Guelpeli et al. 2022] elenca as principais técnicas e abordagens de PLN que são direcionadas para o processamento de texto: Remoção de *Stopwords* (Filtragem de Palavras de Parada); *TF-IDF* (Frequência de Termo - Frequência de Documento Inverso); *Latent Semantic Analysis* (LSA) (Análise Semântica Latente); *N-grams*; *Segmentation* (Segmentação de texto em frases); *Tokenization* (Segmentação de texto em palavras); *Stemming* (Lematização e radicalização); *Part-of-Speech (POS) Tagging* (Etiquetagem morfossintática); e Etiquetagem do Gênero Textual.

Com o objetivo de comparar diferentes técnicas de Inteligência Artificial (IA) na classificação de elementos coesivos, podemos treinar o modelo adotando classificadores como: Árvores de Decisão, Florestas Aleatórias, *Gradiente Boosting*, *Ada Boost*, *Stochastic Gradiente Descent* (SGD), *Support Vector Machines* (SVM), Rede Neural, entre outros. Após a conclusão da RSL, identificaremos de forma mais profunda as estratégias mais utilizadas relacionadas ao uso das técnicas na automatização de redações no contexto da educação básica.

2.5. Avaliação dos resultados

A avaliação dos resultados é um importante passo para o desenvolvimento de qualquer método de classificação. Para aplicar nessa pesquisa, observa-se as seguintes métricas que podem ser utilizadas: (i) **Acurácia:** mede a performance geral do modelo considerando quantas classificações o algoritmo classificou corretamente em relação a todas as

classificações; (ii) **Correlação de *Pearson:*** mede o grau da correlação entre os resultados do algoritmo e os resultados dos avaliadores humanos; (iii) **Precisão:** mede a proporção dos resultados classificados como positivo pelo algoritmo e pelo avaliador humano em relação aos resultados classificados como positivo pelo algoritmo; (iv) ***Recall:*** mede a proporção dos resultados classificados como positivo pelo algoritmo e pelo avaliador humano em relação aos resultados classificados como positivo pelo avaliador humano; (v) ***F1-Score:*** afere a média harmônica entre precisão e a cobertura; (vi) **Erro Médio Absoluto:** afere a média das diferenças entre as correções dos avaliadores humanos e as geradas pelo algoritmo.

A fim de avaliar a eficiência do algoritmo nos experimentos, as métricas listadas podem ser utilizadas em diferentes classificadores. Os resultados dos experimentos também serão comparados aos realizados por analisadores humanos, visando avaliar a eficiência do modelo nos diversos cenários de testes. Dessa forma, o algoritmo proposto será aplicado, de forma experimental, nas avaliações formativas de escrita realizadas em escolas inseridas no programa Brasil na Escola do MEC.

3. Resultados Esperados

Os principais resultados esperados da pesquisa são descritos a seguir:

- Compreender o estado da arte sobre o uso de técnicas de processamento de linguagem natural aplicado no contexto da educação básica brasileira, contribuindo de forma relevante para a comunidade científica;
- Processar um algoritmo eficiente para análise de coesão em redações do segundo segmento do ensino fundamental brasileiro, disponibilizando-o também para futuros pesquisadores;
- Analisar os resultados em um experimento de caso real e verificar a compatibilidade das notas dos elementos de coesão atribuídas pelos professores;
- Disponibilizar um *Dashboard* para subsidiar os professores no ensino de elementos de coesão.

Espera-se com estes resultados a otimização do trabalho do professor, bem como uma contribuição de caráter social por meio da diminuição das desigualdades de aprendizagem de escrita de alunos dos anos finais do ensino fundamental de escolas brasileiras.

4. Considerações Finais

Nessa pesquisa, busca-se automatizar o processo de correção de elementos coesivos em redações de alunos do ensino fundamental. Para isso, a utilização de técnicas de PLN visa apoiar os professores quanto à classificação dos níveis de uso dos elementos de coesão nas redações realizadas durante a formação do estudante. A pesquisa experimental pretende ser de grande valia para os professores de escolas brasileiras de anos finais do ensino fundamental inseridas no programa Brasil na Escola do MEC. Assim, o cumprimento dos objetivos propostos pretende gerar contribuições na perspectiva social e acadêmica.

Este trabalho é desenvolvido no programa de Doutorado Profissional em Engenharia de Software da CESAR School. O ingresso no curso foi em setembro de 2020 na área de pesquisa de *Learning Analytics*. Em Março de 2022, integramos a equipe de pesquisadores do programa Brasil na Escola do MEC em parceria com a Universidade Federal de Alagoas (UFAL). A defesa da tese é prevista para até Agosto de 2024.

Referências

Antunes, I. (2005). *Lutar com palavras: coesão e coerência*. Parábola.

Azmi, A. M., Al-Jouie, M. F., and Hussain, M. (2019). Aaee–automated evaluation of students' essays in arabic language. *Information Processing & Management*.

Cohen, Y., Ben-Simon, A., and Hovav, M. (2003). The effect of specific language features on the complexity of systems for automated essay scoring.

Correnti, R., Matsumura, L. C., Wang, E., Litman, D., Rahimi, Z., and Kisa, Z. (2020). Automated scoring of students' use of text evidence in writing. *Reading Research Quarterly*.

Ferreira Mello, R., André, M., Pinheiro, A., Costa, E., and Romero, C. (2019). Text mining in education. *WIREs Data Mining and Knowledge Discovery*.

Graham, S. (2019). Changing how writing is taught. *Review of Research in Education*.

Grama, D. F. (2022). Elementos coesivos do português brasileiro em córpus de redações nos moldes do enem: um estudo para a elaboração da cotex.

Guelpeli, M. V. C., Fonseca, C. A., and de Souza, R. S. (2022). Representação dos dados estruturados do gênero textual como técnica para o processamento automático de texto.

Halliday, M. A. and Hasan, R. (1995). Cohesion in english.

Ifenthaler, D. (2022). Automated essay scoring systems. In *Handbook of Open, Distance and Digital Education*.

Koch, I. G. V. (1989). *A coesão textual*. Contexto São Paulo.

Lima, F., Haendchen Filho, A., Prado, H., and Ferneda, E. (2018). Automatic evaluation of textual cohesion in essays. In *19th International Conference on Computational Linguistics and Intelligent Text Processing*.

Östling, R., Smolentzov, A., Hinnerich, B. T., and Höglin, E. (2013). Automated essay scoring for swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*.

Passero, G., Ferreira, R., and Dazzi, R. L. (2019). Off-topic essay detection: A comparative study on the portuguese language. *Revista Brasileira de Informática na Educação*.

Pirnay-Dummer, P. and Ifenthaler, D. (2011). Text-guided automated self assessment. a graph-based approach to help learners with ongoing writing. In *Multiple perspectives on problem solving and learning in the digital age*.

Ramesh, D. and Sanampudi, S. K. (2022). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*.

Tian, Y., Kim, M., Crossley, S., and Wan, Q. (2021). Cohesive devices as an indicator of l2 students' writing fluency. *Reading and Writing*.

UNESCO (2011). Everyone has the right to education.

Zupanc, K. and Bosnić, Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems*.