Uma Proposta de Ferramenta para Correção de Textos Descritivos Formais na Língua Portuguesa

Aluizio Haendchen Filho, Vinícius Antònio Cavaco, Jonathan Nau

Núcleo de Inteligência Artificial e Sistemas Inteligentes (NIASI)
Centro Universitário de Brusque (UNIFEBE) – Brusque, SC – Brasil

{aluizio.h.filho,vinicavaco3,jonathan.naau}@gmail.com

Abstract*. Current text editors provide features for identifying and correcting spelling errors as well as some grammatical errors, but they have several limitations. In a sample analysis of 2123 essays, commercial text editors did not detect more than 60% of the errors pointed out by human evaluators. This paper presents a proposal for a text correction application that solves most of the limitations presented by commercial editors.*

1. Introdução

Os atuais processadores de texto, tais como o Word da Microsoft, LibreOffice, dentre outros, oferecem recursos para identificação e correção de erros ortográficos, e de alguns erros gramaticais, mas possuem várias limitações. Para avaliar a eficiência da identificação de erros realizada pelos atuais processadores, desenvolvemos um algoritmo para extrair os erros apontados pelos revisores das redações do site UOL Educação [UOL Educação 2017]. Verificou-se que mais de 60% dos erros apontados pelos avaliadores humanos não são detectados pelos corretores incorporados aos processadores de texto.

Nosso núcleo de pesquisa, o NIASI, desenvolve trabalhos nas áreas de Inteligência Artificial, com aprendizado de máquina, e Processamento de Linguagem Natural. Um dos nossos objetivos é desenvolver ferramentas para automatizar a correção de textos.

Considerando os problemas encontrados nos processadores comerciais de texto, e considerando que existem ferramentas *open-source* com pontos positivos de destaque, escolhemos duas delas para apoiar o desenvolvimento de uma solução. A ferramenta CoGrOO [Colen 2013] se destaca na correção de erros de concordância de gênero e número, e a ferramenta LanguageTool [Miłkowski 2010] mostrou ser mais eficiente na detecção de erros gramaticais e ortográficos. Este trabalho apresenta uma proposta de solução para correção de textos que visa resolver as limitações apresentadas pelos processadores comerciais.

2. Abordagem Utilizada

Para o desenvolvimento da solução, aproveitamos os códigos e regras disponibilizados pelas ferramentas *open-source* citadas, e criamos mais de 200 novas regras, que abrangem concordância, pontuação, crase, palavras homônimas e parônimas, dentre outros não detectados pelos editores comerciais. Foi ainda incorporado um módulo para identificação de desvios de linguagem [Leal 2012], desenvolvido em nosso centro de pesquisa [Nau et al. 2017]. A abordagem utilizada para o desenvolvimento da solução compreendeu os seguintes passos:

1. **Detecção das limitações dos atuais editores**: desenvolvemos um algoritmo para extrair os erros ortográficos e gramaticais apontados pelos revisores das redações do site UOL. As mesmas redações foram submetidas ao CoGrOO para verificar quais erros a ferramenta consegue detectar. Num corpus de 2123 redações, a quantidade de erros detectados pelos editores comerciais (CoGrOO como base) alcançou 6.868 erros, ao passo que os revisores humanos detectaram 18.923 erros, ou seja, 63,7% dos erros não foram detectados pela ferramenta.
2. **Projeto da ferramenta**: após a análise dos erros detectados, foi incorporado ao projeto os código fonte *open-source* das ferramentas, selecionando as funcionalidade de cada uma com melhor resultado na detecção de erros.
3. **Implementação**: após a análise dos erros não detectados, foi implementado um protótipo, cujas principais funcionalidades estão descritas a seguir.

3. Funcionalidades

O Diagrama de Atividades, apresentado na Figura 1, mostra as principais funcionalidades do sistema, a sua interação com as ferramentas CoGrOO, LanguageTool, e com o catálogo de desvios de linguagem. As atividades iniciam e são concluídas pelo Usuário.

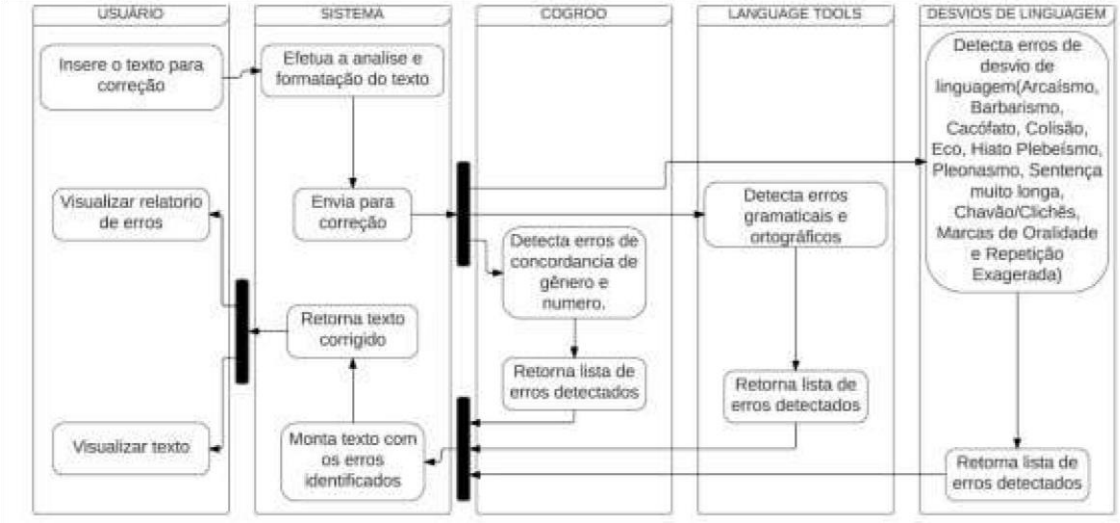

Figura 1. Diagrama de Atividades da ferramenta implementada.

Após inserir o texto para a correção, o sistema armazena o texto e envia para correção. Os erros são identificados pelo sistema de regras obtido das ferramentas CoGrOO e LanguageTool, e pelas novas regras incorporadas ao código *open-source*. A seguir, o texto é submetido ao módulo de detecção de vícios e desvios de linguagem, que aponta os erros detectados. O sistema então marca com cores os erros detectados no texto, e apresenta ao usuário. O usuário então pode aceitar as sugestões de correção disponibilizadas pelo sistema, ou ignorá-las.

4. Resultados

Após a incorporação das novas regras e inclusão de vícios de linguagem, a solução proposta apresentou os seguintes resultados, conforme mostra a Tabela 1.

Tabela 1. Comparativo da quantidade de erros detectados pela ferramenta

Quantidade de Redações	Quantidade de erros detectados pelos revisores	Quantidade de erros detectados por editores comuns	Quantidade de erros detectados pela ferramenta
2123	18923	6868	21445

Como pode ser visto, a ferramenta detectou 21.445 erros, enquanto os editores comerciais detectaram somente 6.868 erros. Neste mesmo corpus, os revisores humanos detectaram 18.923 erros. Portanto, a ferramenta detectou uma grande quantidade de erros não identificada pelos editores comerciais, principalmente erros gramaticais tais como concordância, pontuação. Detectou ainda uma quantidade maior de erros do que aqueles identificados pelos revisores humanos. Dentre erros detectados pela ferramenta e não apontados pelos avaliadores humanos, verificamos que, em sua maioria, são erros de vícios de linguagem, tais como eco (928 erros), e hiato (636), marcas de oralidade (303), e colisão (273 erros), dentre outros.

5. Conclusão

Na etapa inicial de testes da solução desenvolvida, verificamos que ela se mostrou apta a encontrar erros ortográficos, gramaticais, além de vícios de linguagem não detectados pelos processadores comerciais de texto. A abordagem é inédita, e demandou considerável pesquisa para identificar tipos de erro não detectados, avaliar as soluções existentes, criar novas regras e implementar uma solução eficaz. A ferramenta pode ser utilizada para auxiliar a correção de textos descritivos formais, tais como TCCs, artigos, trabalhos acadêmicos e textos em geral. Além disso, a solução poderá também ser utilizada em contextos tais como correção automatizada de redações em concursos e para a avaliação da Competência 1 do ENEM (norma padrão da língua portuguesa).

Agradecimento

Agradecemos à Secretaria de Educação do Estado de Santa Catarina – Programa de Bolsas de Pesquisa Artigo 170 – pelo auxílio financeiro a este projeto de pesquisa.

Referências

Colen, W. (2013) Aprimorando o corretor gramatical CoGrOO. São Paulo: USP.

Leal, E. S. (2012) Vícios de Linguagem e Idiotismos: A Fala Como Unidade de Estudos nas Gramáticas Normativas Brasileiras em Língua Portuguesa entre 1881-1890. Páginas 255-265. Revista Uniletras. 2012.

Miłkowski, M. (2010) Developing an open-source, rule-based proofreading tool. Journal of Software Practice and Experience, New York, pages 543-566.

Nau, J., Haendchen Filho, A., Passero, G., and Cavaco, V. (2017) Uma Ferramenta para Identificar Desvios de Linguagem na Língua Portuguesa. XI Brazilian Symposium in Information and Human Language Technology and Collocated Events. Proceedings of the Conference October 2 to 5, 2017. Uberlândia, Brazil.

UOL Educação (2017) "Banco de Redações". Obtido em: https://educacao.uol.com.br/bancoderedacoes/, Maio, 2017.