Revisor Gramatical

Italo Teixeira da Silveira[1], **Murillo Aleixo Mota**[1], **Rogério Farias Otto**[1]

[1]Universidade Federal de Pelotas (UFPel)

{itdsilveira,mamota,rfotto}@inf.ufpel.edu.br,

Abstract. *The project consists of a grammar checker in which the user will write a text as an input that will be passed through a spell check and then a grammar check that will result in the corrected text, thus requiring only a basic level of portuguese.*

Resumo. *O projeto consiste em um corretor gramatical onde o usuário vai escrever um texto como uma entrada e a partir disso ocorrerá a correção ortográfica e uma revisão gramatical que resultará em um texto corrigido, sendo assim precisando apenas de um nível básico da língua portuguesa.*

1. Introdução

Nosso trabalho se trata de um revisor gramatical para um editor de texto qualquer, onde a entrada será textos com erros gramaticais que o editor irá detectar e a saída dele será os textos corrigidos ortograficamente e com erros de concordância indicados. Nosso público alvo abrange qualquer pessoa que queira escrever textos na linguagem portuguesa, onde este usuário deverá ter um nível de conhecimento extremamente básico da língua portuguesa para que o fornecimento de uma entrada ao programa seja possível.

2. Descrição do Dataset, Técnicas e Projeto Proposto

2.1. Dataset

O dataset foi criado utilizando como base o dataset criado por Murilo Gazola, Sidney Evaldo Leal, Sandra Maria Aluisio. O nome do córpus é: "Córpus de Complexidade Textual para Estágios Escolares do Sistema Educacional Brasileiro"[1], onde o córpus inclui trechos de: livros-textos cuja lista completa é apresentada abaixo, notícias da Seção Para Seu Filho Ler (PSFL) do jornal Zero Hora que apresenta algumas notícias sobre o mesmo córpus do jornal do Zero Hora, mas escritas para crianças de 8 a 11 anos de idade , Exames do SAEB , Livros Digitais do Wikilivros em Português, Exames do Enem dos anos 2015, 2016 e 2017. Todo o material em português foi disponibilizado para avaliar a tarefa de complexidade textual (readability).

Lista completa dos livros Didáticos e suas fontes originais:

Título - Assunto	Editora	Série
Marcha criança - lingua portuguesa -	Scipione didáticos	3º ano
Tudo é linguagem - literatura	Ática	3º ano
Projeto Porta Aberta - lingua portuguesa	FTD	3º ano
Projeto Ápis - lingua portuguesa	Atica	3º ano
Português - lingua portuguesa	Atual	3º até 7º ano
Presente - lingua portuguesa	Moderna	4º ano
Buriti - lingua portuguesa	Leya Escolar	5º ano
Porta Aberta - lingua portuguesa	FTD	5º ano
Mundo Amigo - lingua portuguesa	SM Brasil	5º ano
Nos Dias de Hoje - lingua portuguesa	Leya Escolar	6º ano
Projeto Teláris - lingua portuguesa	Ática	6º e 7º ano
Varios titulos - Varios tópicos	CNEC — Educação	Ensino fundamental

2.2. Técnicas

Primeiramente o texto cru passa pelo algoritmo Spell Checker, criado por Andrés Segura Tinoco que é um algoritmo baseado no teorema de Bayes que faz a correção ortográfica do texto. O Teorema de Bayes é uma fórmula matemática usada para o cálculo da probabilidade de um evento dado que outro evento já ocorreu, o que é chamado de probabilidade condicional. A respeito do teorema, precisa-se ter informações anteriores, ou seja, precisa-se saber que um determinado evento já ocorreu e qual a probabilidade deste evento.

Logo após o processo de correção ortográfica, utiliza-se a biblioteca Spacy para analizar a gramática. Utilizamos da biblioteca para comparar se duas palavras são do mesmo gênero e pluralidade, para então mostrar para o usuário quais são os erros de concordância.

3. Análise dos resultados obtidos

Os resultados obtidos foram satisfatórios, o programa consegue corrigir a maioria das frases que lhe é passada e aponta os erros gramaticais, talvez o programa fosse mais preciso se o dataset utilizado para a correção ortográfica fosse maior, pois tem algumas palavras que tem baixa aparição no dataset que atrapalha na hora da correção. A lógica de verificação gramatical não é muito robusta, então as vezes alguns erros passam batido pelo programa.

Por exemplo, a entrada do programa é 'qur abaçar um árvre?', o resultado esperado da correção é 'quer abraçar uma árvore?', porém, o programa corrige somente as palavras que estão mal escritas, logo o resultado será 'que abraçar um árvore?', nota-se que 'qur' vira 'que' pois a probabilidade é maior que 'quer'. No proximo passo analisamos o gênero das palavras, e o programa aponta que 'um' não combina com 'árvore' pois deveria ser 'uma'.

4. Considerações finais

Deste modo, pelo fato do dataset utilizado não apresentar uma quantidade vasta e nem variada de palavras o suficiente, ocorrem vezes em que não funciona da maneira esperada, porém, na maioria das vezes o algoritmo mostra onde o erro existe e corrige-o, cumprindo assim, a proposta do projeto com o auxílio do Spell Checker e do Spacy.

References

[1] Sandra Maria Aluisio Murilo Gazzola, Sidney Evaldo Leal. Predição da complexidade textual de recursos educacionais abertos em português. In *Proceedings of the Brazilian Symposium in Information and Human Language Technology*, 2019.