GRAMÁTICAS LOCAIS PARA VERIFICAÇÃO GRAMATICAL DO PORTUGUÊS

Caio Pereira Lapa, João Victor Mascarenhas de Faria Santos, Juliana Pinheiro Campos Pirovani.

Universidade Federal do Espírito Santo/Centro de Ciências Agrárias, Alto Universitário, s/n, Caixa Postal 16, Guararema - 29500-000 - Alegre-ES, Brasil, caio.lapa@edu.ufes.br, joao.vmf.santos@edu.ufes.br, juliana.campos@ufes.br·

Resumo

O reconhecimento de erros gramaticais e ortográficos em textos não estruturados é um desafio no Processamento de Linguagem Natural (PLN), especialmente na língua portuguesa, dada sua complexidade. Identificar erros como o uso correto da crase e da vírgula, é crucial para assistentes virtuais, traduções automáticas e para a revisão de documentos. Corretores gramaticais com abordagens linguísticas têm um papel fundamental nesse processo. Este trabalho tem como objetivo investigar a aplicação das Gramáticas Locais na verificação gramatical de sentenças em português. As Gramáticas Locais (GLs), baseadas em autómatos de estados finitos, capturam expressões com características sintáticas ou semânticas em comum. Pretende-se analisar a viabilidade de incorporar essas GLs em ferramentas já estabelecidas, como o CoGrOO (2023), buscando aprimorar sua capacidade de verificação. Foram feitas 14 GLs para reconhecimento de erros de vírgula e crase em sentenças. Também foi investigada a possibilidade de integração dessas GLs em ferramentas existentes, como o CoGrOO, para aprimorar suas capacidades de verificação gramatical.

Palavras-chave: Gramáticas Locais. Verificação Gramatical. Língua Portuguesa. Processamento de Linguagem Natural. CoGrOO.

Área do Conhecimento: Ciências Exatas e da Terra, Ciência da Computação.

Introdução

Uma quantidade significativa de informações está contida em textos não estruturados, redigidos de forma livre. Para aprimorar a precisão na correção de diversos tipos de textos, o reconhecimento de erros gramaticais e ortográficos se mostra crucial. Esta tarefa, inserida no domínio do Processamento de Linguagem Natural (PLN), visa identificar automaticamente erros em textos diversos. No entanto, reconhecer esses erros é um desafio devido à variabilidade de sentidos e contextos que uma alteração sutil na escrita pode ocasionar. Além disso, a língua portuguesa apresenta características estruturais complexas que podem representar desafios para os sistemas de PLN, especialmente no que diz respeito à detecção de erros relacionados ao uso adequado da crase e da vírgula.

O reconhecimento desses erros é influenciado tanto pelo idioma quanto pela mensagem que o texto transmite ao leitor. Dependendo da mensagem pretendida, o mesmo texto pode ser estruturado de maneiras muito similares, mas com significados diferentes. Por exemplo, assistentes virtuais precisam identificar corretamente o uso da crase e interpretar adequadamente a pontuação, especialmente a vírgula, já que estas podem alterar o significado das frases. Na tradução automática, a crase impacta diretamente a conversão entre idiomas. Já na análise de sentimentos, a pontuação influencia o tom emocional de um texto. Ferramentas de correção automática também dependem de uma compreensão sólida dessas regras gramaticais, sugerindo o uso correto da crase e ajustando a vírgula em frases compostas. Assim, o domínio da crase e da vírgula é essencial para garantir a eficácia de assistentes virtuais, sistemas de tradução automática, análise de sentimentos e correção de documentos em processamento de linguagem natural (PLN). Logo, corrigir esses possíveis erros é importante e fundamental para a revisão de documentos, para a melhoria da escrita e comunicação de forma geral, e em tarefas de PLN.

Os corretores gramaticais, também conhecidos como verificadores ou revisores gramaticais, utilizam algoritmos e regras para identificar possíveis problemas ao analisar o texto escrito em busca de erros. Esses corretores podem utilizar diferentes abordagens na sua implementação. Na abordagem

linguística, as regras são criadas manualmente, com base nas regras gramaticais e ortográficas, para detectar erros comuns. Já no aprendizado de máquina, os sistemas aprendem a identificar e corrigir erros com base em um conjunto de dados de treinamento. A abordagem híbrida, por sua vez, combina elementos das duas abordagens anteriores para obter um sistema mais abrangente e eficiente. Neste trabalho, será utilizada apenas a abordagem linguística para verificação gramatical.

Uma maneira de formalizar as regras da abordagem linguística são as Gramáticas Locais (GLs), concebidas por Maurice Gross (GROSS, 1997). Segundo Gross (1999), as Gramáticas Locais são sistemas de estados finitos ou autômatos de estados finitos que representam conjuntos de expressões em uma língua natural. Essas gramáticas são elaboradas manualmente e servem para agrupar ou capturar expressões que compartilham características em comum, sejam elas sintáticas ou semânticas.

Este trabalho tem como objetivo investigar a aplicação das Gramáticas Locais na verificação gramatical de sentenças em língua portuguesa. Pretende-se analisar a viabilidade de incorporar essas gramáticas em ferramentas já estabelecidas, como o CoGrOO (2023), visando aprimorar sua capacidade de verificação.

Metodologia

Inicialmente foi realizada uma revisão da literatura para aprofundar os conhecimentos dos temas relevantes para o contexto deste trabalho, sendo estes: Gramática Local (GROSS, 1999), Unitex (MUNIZ et al., 2005), Processamento de Linguagem Natural (PLN) e verificação gramatical do Português.

Segundo Gross (1999), as Gramáticas Locais são sistemas de estados finitos ou autômatos de estados finitos que representam conjuntos de expressões em uma língua natural. O Unitex é um conjunto de *software* livres para PLN que permite construir GLs e aplicá-las, assim como Cascatas de Transdutores, realizar pré-processamento de textos e aplicação de dicionários (PAUMIER, 2021). A verificação gramatical em Português refere-se ao processo de analisar um texto escrito em Português para identificar e corrigir erros gramaticais.

A seguir, foram identificados corretores gramaticais de código aberto para o Português que poderiam ser utilizados e adaptados neste trabalho: CoGrOO, LanguageTool e o Hunspell. O corretor gramatical escolhido foi o CoGroo (CoGrOO, 2023), já que é uma ferramenta específica para a língua portuguesa. Isso significa que é altamente adaptado às particularidades gramaticais e ortográficas do idioma, proporcionando uma verificação mais precisa e relevante para textos em Português. Além disso, é um projeto de código aberto (seu código fonte está disponível para acesso) que vem sendo bastante utilizado e pode ser acoplado ao LibreOffice (COSTA et al., 2020).

A versão 3.1.0 do CoGrOO foi analisada. É uma ferramenta de análise gramatical desenvolvida para a língua portuguesa que adota uma abordagem híbrida que combina técnicas estatísticas e baseadas em regras (KINOSHITA et al., 2005) para corrigir erros gramaticais. Ele emprega um conjunto de regras linguísticas definidas para identificar e corrigir erros em textos escritos em Português.

Na sequência, uma base de exemplos com sentenças de sites como Nova Escola e Norma Culta foi criada para os testes de vulnerabilidade do CoGrOO durante o período de familiarização com o corretor. A base de exemplos apresenta dois arquivos distintos: um com exemplos de uso correto da crase e um com exemplos de uso correto da vírgula. Cada arquivo apresenta 40 sentenças com diversos usos tanto da crase quanto da vírgula.

Foram realizados testes no CoGrOO com as sentenças da base de exemplos, retirando as crases nas sentenças que contêm crase e retirando as vírgulas das sentenças que contêm vírgula. Após isto, o CoGrOO deveria identificar e apontar os erros nas sentenças para que a correção fosse feita. Após realizar os testes, investigando suas vulnerabilidades e flexibilidades, observou-se que há a necessidade de melhoria nas correções a respeito do uso da crase e da vírgula.

Resultados

Durante os testes realizados, o CoGrOO corrigiu 10 sentenças em um conjunto de 80 exemplos. Dentre as sentenças com erros de crase, foram corrigidas 4 de forma adequada. Já em relação aos erros de vírgula, o CoGrOO corrigiu corretamente 6 sentenças.

Observou-se que o corretor de texto apresenta uma melhor detecção e correção de erros em apenas algumas regras de uso de crase, tais como:

1. Antes de palavras femininas em construções de sentenças com substantivos e adjetivos que requerem a preposição “a” e antes de verbos cuja regência é feita com a preposição “a”:
   a. Errado: É importante obedecer às regras de funcionamento da escola.
   b. Corrigido: É importante obedecer às regras de funcionamento da escola.

2. Antes de nomes de localidades que permitem a anteposição do artigo "a" quando regidos pela preposição "a":
   a. Errado: Vamos a loja.
   b. Corrigido: Vamos à loja.

Por outro lado, em relação às correções relacionadas ao uso da vírgula, o corretor obtém os melhores resultados em outro conjunto restrito de regras, tais como:

1. Isolar expressões intercaladas na oração, como por exemplo, contudo, todavia, além disso, logo e enfim, entre outras, usadas para expressar oposição, explicação, conclusão:
   a. Errado: Será necessário por exemplo um novo computador.
   b. Corrigido: Será necessário, por exemplo, um novo computador.

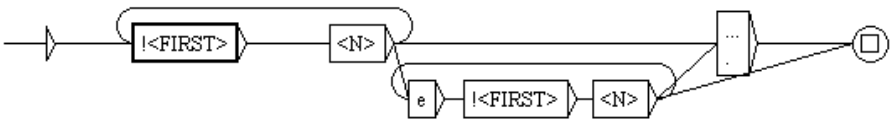
2. Isolar os advérbios "sim" e "não" quando iniciam uma frase que serve de resposta a uma frase anterior:
   a. Errado: Sim vocês podem contar com nossa ajuda.
   b. Corrigido: Sim, vocês podem contar com nossa ajuda.

3. Isolar o nome do lugar na escrita da data, sendo frequentemente empregada em cartas e outras comunicações:
   a. Errado: Ouro preto 31 de março de 2013.
   b. Corrigido: Ouro preto, 31 de março de 2013.

Com o intuito de aprimorar a detecção e correção desses erros, 14 GLs foram desenvolvidas até o momento, visando suprir as limitações do CoGrOO nesses aspectos. Abaixo, seguem 2 exemplos de GLs criadas, ilustradas nas imagens 1 e 2, com suas respectivas regras e exemplos:

1. Separação de elementos coordenados (independentes um do outro na oração):
   - Eu como todo tipo de comida, arroz, feijão, couve, farofa…

Figura 1 - GL que identifica a falta de separação de elementos repetidos por vírgula



Fonte: Criado pelo autor.

Para detectar erros nos exemplos relacionados à Regra 1, utilizou-se !<FIRST> junto com <N> em um loop para identificar um ou mais substantivos que começam com letra minúscula. Aqui está como o processo funciona:

A. <!FIRST>: O operador ! (negação) seguido do código lexical <FIRST> do Unitex é utilizado para excluir substantivos que começam com letra maiúscula, focando assim nos substantivos que estão com a primeira letra minúscula, o que indica um possível erro.

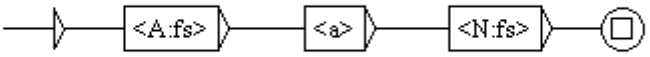
B. <N> em loop: A gramática permite detectar uma sequência de substantivos com a letra inicial minúscula. Isso é útil para lidar com casos onde há mais de um substantivo em sequência.

C. Depois disso, a gramática procura por:
   a. **“...”** ou **“.”**: Reticências ou ponto final são utilizados para identificar o final da sentença.
   b. “e”: A presença da conjunção "e" é identificada para marcar o último elemento coordenado na sentença.

Com essa estrutura, é possível capturar erros relacionados a substantivos que deveriam começar com maiúscula, mas estão em minúscula, considerando também o contexto de uma coordenação no final da sentença.

2. Uso de crase quando o complemento de um nome exige a preposição “a” seguida de um substantivo feminino antecedido de artigo feminino “a”:
   - “Ela se mostrou favorável à medida proposta pela vereadora.”

Figura 2 - GL que identifica falta de crase



Fonte: Criado pelo autor.

Para detectar erros nos exemplos relacionados à Regra 2, a gramática utiliza a tag <A> para identificar adjetivos femininos no singular. Em seguida:

A. A preposição "a" é marcada pela tag <a>.
B. Substantivos femininos no singular são identificados pela tag <N>.

A sentença é considerada correta apenas se o adjetivo feminino, a preposição "a" e o substantivo feminino forem encontrados na estrutura correta e na ordem esperada. Ou seja, a gramática verifica se a sequência segue a estrutura: *adjetivo feminino + preposição “a” + substantivo feminino*. Caso contrário, a sentença é marcada como incorreta.

As 14 GLs criadas foram testadas na mesma base de dados que o CoGroo. O conjunto de GLs identificou 46 erros das 80 sentenças até o momento, apresentando uma melhora na detecção dos erros em relação ao CoGroo.

Discussão

Após a análise dos resultados dos testes realizados, foi verificado que o corretor CoGrOO apresenta deficiências na detecção e correção de sentenças, tornando-se necessário o desenvolvimento de Gramáticas Locais (GLs) para aprimorar a detecção desses erros.

As Gramáticas Locais (GLs) desenvolvidas demonstraram um avanço significativo no reconhecimento de erros gramaticais, apresentando uma melhoria expressiva de 45% na detecção de sentenças incorretas em comparação ao CoGrOO. Esse salto na precisão é notável, especialmente em áreas críticas como a crase e o uso da vírgula, com 20 sentenças adicionais identificadas para crase e 26 para vírgula.

A superioridade das GLs mostra o seu potencial, mas, ainda mais importante, ressalta a oportunidade de aprimoramento contínuo. Com ajustes e testes em diferentes tipos textuais, essas GLs podem elevar ainda mais a qualidade e a abrangência da detecção de erros.

Dentre os erros que não foram ou talvez não sejam possíveis de serem identificados, estão regras de omissão de palavras, como por exemplo a regra de crase:

1. É necessário crase antes de palavra masculina, caso haja uma palavra feminina implícita:
   a. Necessitamos, com urgência, ir à abastecimentos. (Isto é: Necessitamos, com urgência, ir à central de abastecimento.)

Conclusão

Neste trabalho, foram desenvolvidas 14 Gramáticas Locais (GLs) com o objetivo de aprimorar a detecção de erros de crase e vírgula no português. As GLs demonstraram resultados expressivos, superando a taxa de acerto do CoGrOO ao serem aplicadas às mesmas sentenças da base de dados. A eficácia das GLs na identificação de erros específicos destaca uma melhoria significativa em relação às soluções já existentes, indicando seu grande potencial para elevar o desempenho do CoGrOO quando forem integradas.

Como próximas etapas, o foco será no aperfeiçoamento contínuo dessas GLs, garantindo sua precisão ao testá-las em outros tipos de textos e contextos linguísticos. Além disso, a integração dessas gramáticas ao CoGrOO, iniciada mas ainda não concluída, permanece uma prioridade a ser alcançada para consolidar esses avanços.

Referências

Academia. 2024. Acesso em: 05/02/24. Disponível em: <https://www.academia.org.br/>

COSTA, Luciana; DE OLIVEIRA, Elaine Harada Teixeira; JÚNIOR, Alberto Castro. **Corretor Automático de Redações em Língua Portuguesa: um mapeamento sistemático de literatura.** In: Anais do XXXI Simpósio Btasileiro de Informática na Educação. SBC, 2020. p. 1403-1412.

GROSS, M. **A Bootstrap Method for Constructing Local Grammars.** In: BOKAN, N. (Ed.). Proceedings of the Symposium on Contemporary Mathematics. Belgrado, Sérvia: University of Belgrad, 1999. p. 229– 250.

GROSS, M. **The Construction of Local Grammars.** In ROCHE, E.; SCHABÈS, Y. (eds.). Finite-state language processing, Language, Speech, and Communication, Cambridge, Mass**.**, MIT Press, p. 329–354, 1997.

KINOSHITA, Jorge.; SALVADOR, Laís do Nascimento.; MENEZES,Carlos E. Dantas. CoGrOO – **Um Corretor Gramatical para a língua portuguesa, acoplável ao OpenOffice.** São Paulo, Brazil, 2005. Disponível em: <http://valinhos.ime.usp.br:55080/cogroo/sites/ime.usp.br.cogroo/files/CLEI_COGROO_2005.pdf>

MUNIZ, M. C. M.; NUNES, M. G. V.; LAPORTE, E. **UNITEX-PB, a set of flexible language resources for Brazilian Portuguese.** Workshop on Technology on Information and Human Language (TIL), 2005, São Leopoldo, Brazil. pp.2059-2068.

PAUMIER, S. **Unitex 3.3 User Manual.** 2021. 393 p. Disponível em: <https://unitexgramlab.org/releases/3.3/man/Unitex-GramLab-3.3-usermanual-en.pdf>.

The News. 2024. Acesso em: 03/02/24. Disponível em: <https://thenewscc.beehiiv.com/>.

Agradecimentos

Agradeço ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) pelo financiamento da bolsa de estudos durante o desenvolvimento deste trabalho.