GRAMÁTICAS LOCAIS PARA O RECONHECIMENTO DE ENTIDADES NOMEADAS EM BULAS DE MEDICAMENTOS

Thiago Tonelli da Silva[1], Kailany Alves Silva[1], Juliana Pinheiro Campos Pirovani[1], Cristiano da Silveira Colombo[2].

[1]Universidade Federal do Espírito Santo/Departamento de Computação, Alto Universitário, S/N, Guararema - 29500-000 - Alegre-ES, Brasil, thiago.silva.66@edu.ufes.br, kailany.silva@edu.ufes.br, juliana.campos@ufes.br.

[2]Instituto Federal do Espírito Santo/Departamento de Informática, Rodovia ES-482, S/N, Fazenda Morro Grande - 29311-970 - Cachoeiro de Itapemirim-ES, Brasil, cristianos@ifes.edu.br.

Resumo

Este artigo visa construir Gramáticas Locais (GLs) para o Reconhecimento de Entidades Nomeadas (REN) em bulas de medicamentos, devido à necessidade de identificá-las automaticamente para agilizar a compreensão de informações essenciais e elucidar dúvidas. Para isso, foi proposto o uso do Processamento de Linguagem Natural (PLN) para extrair dados importantes dos textos. Sendo assim, foram desenvolvidas e ajustadas Gramáticas Locais com base em padrões identificados nos textos, que permitiram reconhecer automaticamente entidades, como nomes de excipientes, princípios ativos, efeitos colaterais e enfermidades. Além disso, o desempenho das GLs foi avaliado através de métricas, calculadas pelo *script* de avaliação do segundo HAREM[1]. Os resultados se mostraram promissores inicialmente, como *precisão: 74%* e *abrangência: 32%.* No entanto, requer mais tempo e atenção há alguns pontos, como a criação de novas GLs para melhorar a abrangência das Entidades Nomeadas (ENs) reconhecidas, e a adaptação das já existentes para melhorar a precisão.

Palavras-chave: Processamento de Linguagem Natural. Bulas de medicamentos. Reconhecimento de Entidades Nomeadas. Gramáticas Locais.

Área do Conhecimento: Ciências Exatas e da Terra - Ciência da Computação

Introdução

O PLN é uma subárea da Ciência da Computação e da Linguística dedicada ao estudo da geração e compreensão da linguagem natural. O REN é uma tarefa que visa identificar e classificar automaticamente entidades, como nomes de pessoas, lugares, organizações, dentre outras, em textos de escrita livre (Pirovani, 2019). É uma tarefa importante para o PLN, pois identificar e categorizar entidades contribui para uma melhor compreensão do contexto, facilitando a interpretação do significado pelo sistema.

O HAREM[1], que foi uma avaliação conjunta para o REN em português e os corpora anotados utilizados no Primeiro e no Segundo HAREM[1], conhecidos como Coleções Douradas (GCs), têm sido empregados como referência padrão-ouro para os sistemas de REN em português (Pirovani; Oliveira, 2021). No trabalho de Pirovani (2019) foram construídas Gramáticas Locais (GLs) para as 10 categorias de EN do HAREM[1], que são *PESSOA*, *ORGANIZACAO*, *LOCAL*, *TEMPO*, *VALOR*, *ABSTRACAO*, *ACONTECIMENTO*, *COISA*, *OBRA*, e *OUTRO*. Essa Gramática Local (GL) foi utilizada em sua abordagem híbrida CRF+LG.

As GLs são autômatos de estados finitos que representam um conjunto de expressões de uma linguagem natural (Gross, 1997). Com elas é possível capturar o contexto em que as ENs aparecem.

Bulas farmacêuticas são documentos que auxiliam pacientes e médicos informando-os sobre a administração, composição, efeitos colaterais e interações com outros medicamentos. O REN é essencial para extrair informações relevantes que facilitam o entendimento desses documentos, visto que, frequentemente, as bulas não fornecem condições claras para o paciente entender como o medicamento poderá ajudá-lo (Silva et al., 2000).

O trabalho de Colombo e Oliveira (2022) teve como objetivo identificar entidades nomeadas em bulas médicas e relatos de casos clínicos. Ele utilizou a abordagem híbrida CRF+LG (Pirovani, 2019)

[1] https://www.linguateca.pt/HAREM/

e observou que algumas entidades do contexto médico, como *omeprazol* (substância) e *taquicardia (*sintoma), foram classificadas nas categorias *COISA* e *ABSTRACAO*, respectivamente. Isso sem ter uma GL específica para esse contexto.

Assim, este trabalho possui como objetivo construir GLs específicas para o REN em bulas de medicamentos, e classificá-las de acordo com as categorias definidas.

Metodologia

As abordagens utilizadas no desenvolvimento de sistemas de REN são: linguística, aprendizado de máquina ou híbrida (Pirovani, 2019). A abordagem adotada neste trabalho será a linguística, onde regras são descritas manualmente para identificar um contexto em que uma Entidade Nomeada (EN) aparece. Contudo, há dificuldades, pois contextos iguais, não necessariamente indicam uma EN corretamente. A Tabela 1 demonstra um exemplo de mesmo contexto e o reconhecimento errôneo (falso-positivo), *dosagem alta* não é uma substância, logo não deve ser anotada.

Tabela 1- Exemplos de contextos.

Contextos a esquerda	EN	Contextos a direita	Sentença
Tratamento com	amoxicilina	(Qualquer coisa)	"Tratamento com amoxicilina"
Tratamento com	dosagem alta	(Qualquer coisa)	"Tratamento com dosagem alta"

Fonte: o autor.

Para observar os contextos tanto à esquerda quanto à direita, visando captar padrões e palavras que podem indicar, de alguma forma, se o que aparece antes ou depois é uma EN, foi realizado um estudo linguístico em alguns documentos pertencentes ao *corpus* construído e anotado manualmente, com a ferramenta Etiquet(H)arem[2], por Colombo e Oliveira (2022). Para este trabalho, foram selecionados 19 documentos desse *corpus*, incluindo 10 relatos médicos da SciELO[3] e 9 bulas de medicamentos para o trato digestivo: *Amoxicilina, Cloridrato de Ranitidina, Esogastro, Gastrium, Label, Iniparet, Laflugi, Omepramix, Pyloripac e Ziprol.*

As GLs foram criadas para reconhecer as categorias *COISA* e *ABSTRACAO*, que indicam substâncias e sintomas, respectivamente. A categoria *COISA* classifica substâncias em dois tipos: *Principio* e *Componente*. O *Principio* diz respeito aos princípios ativos, como *amoxicilina,* e *Componente* refere-se a substâncias não ativas que são adicionadas à fórmula para auxiliar na estabilidade, solubilidade, absorção e palatabilidade do medicamento.

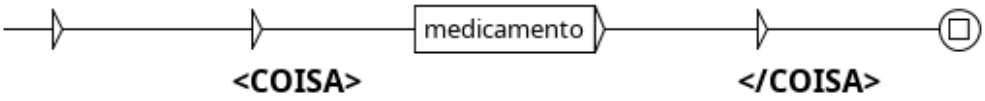
As ENs reconhecidas serão anotadas, ou seja, demarcadas entre *tags*, utilizando o padrão *<CATEGORIA_TIPO>EN</CATEGORIA_TIPO>*, por exemplo *<COISA_PRINCIPIO>amoxicilina </COISA_PRINCIPIO>*.

Para desenvolver as GLs foi utilizado o Unitex[3], um conjunto de *software* livres para PLN, que permite, com base em contextos, identificar as entidades e realizar as anotações. Para melhor entendimento, a Figura 1 apresenta um exemplo de uma GL que reconhece a palavra *medicamento* e anota. Se a palavra for reconhecida, ocorre a transição de estados até o estado final e a palavra é anotada.

Figura 1 - GL que reconhece *medicamento*

[2] http://www.linguateca.pt/poloCoimbra/recursos/etiquetharem.zip

[3] https://unitexgramlab.org/pt



Fonte: o autor.

Por fim, com base nas regras observadas, foram implementados testes e criadas GLs. O desempenho das GLs construídas foi avaliado comparando os resultados da sua identificação com os das bulas anotadas manualmente. Dito isso, foram utilizados os scripts de avaliação do segundo HAREM, que computam métricas comumente utilizadas para REN (precisão, abrangência e medida-F) (Mota e Santos, 2008, apud Pirovani, 2019).

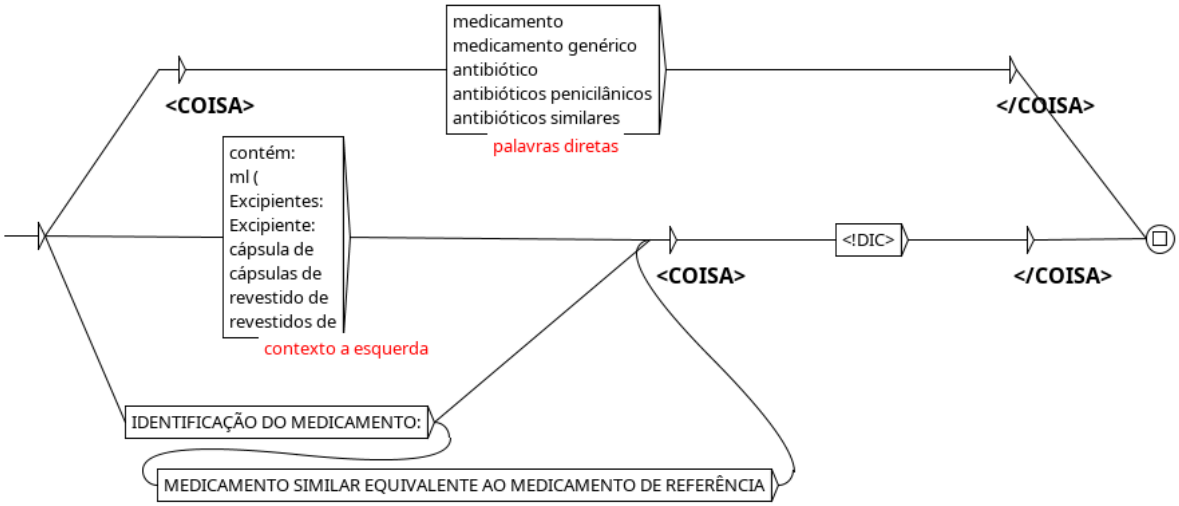
> A precisão representa a quantidade de acertos no total de ENs identificadas. Quanto maior a precisão, menor o número de ENs identificadas de forma errada (falso-positivos). A abrangência representa a quantidade de acertos no total de ENs existentes. Quanto maior, menor a quantidade de ENs não identificadas (falso-negativos). A medida-F é uma média harmônica das outras duas (Pirovani, 2019, p.47).

Resultados

Apresentamos aqui algumas das GLs construídas neste trabalho até o momento. Primeiramente, foi criada a GL *Coisa.grf* mostrada na Figura 2, utilizada para identificar a categoria *COISA,* os tipos não foram considerados. Foram inseridas palavras comuns e recorrentes, para anotá-las sem análise do contexto.

Alguns contextos à esquerda foram considerados, e junto deles foi utilizado o *<!DIC>, tag* lexical utilizada pelo Unitex para palavras que não foram reconhecidas pelo dicionário do Unitex. Isso foi feito pois grande parte das substâncias não estão contidas no dicionário, o que auxilia no reconhecimento e assertividade dessas entidades.

Figura 2 - Coisa.grf



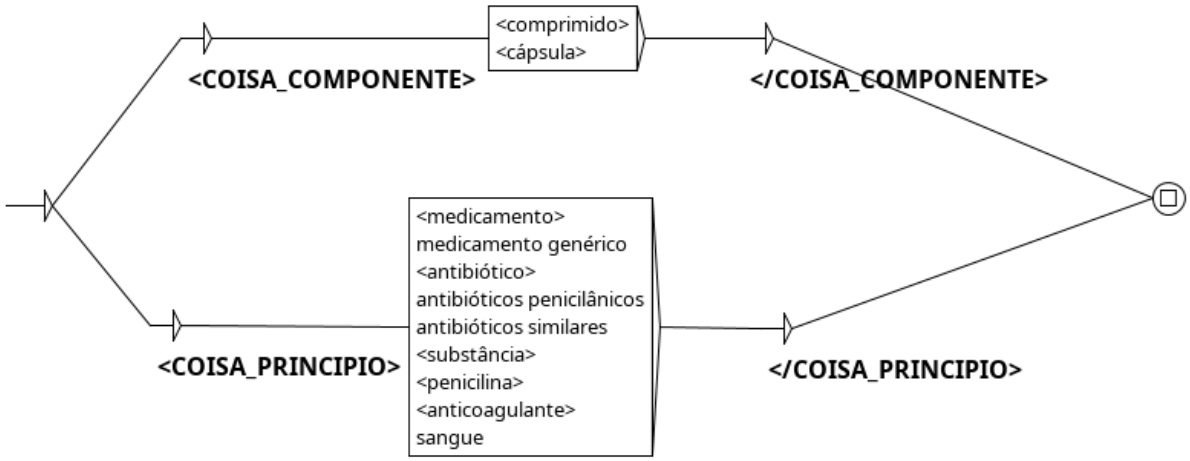
Fonte: o autor.

Após isso, para melhor organização e compreensão, foram construídas GLs individuais para contextos à esquerda e palavras sem contexto. Posteriormente, novos padrões observados eram inseridos ou adaptados nas GLs.

Durante a análise do *corpus* e testes, notou-se que, palavras comumente utilizadas em bulas, poderiam ser anotadas independente do contexto e que essas abrangem significativamente o domínio.

A GL da Figura 3 *palavrasDiretas.grf,* demonstra alguns exemplos de palavras, e anota-as de acordo com o seu tipo. Algumas palavras estão entre < >, isso para poder reconhecê-las no singular e também no plural.

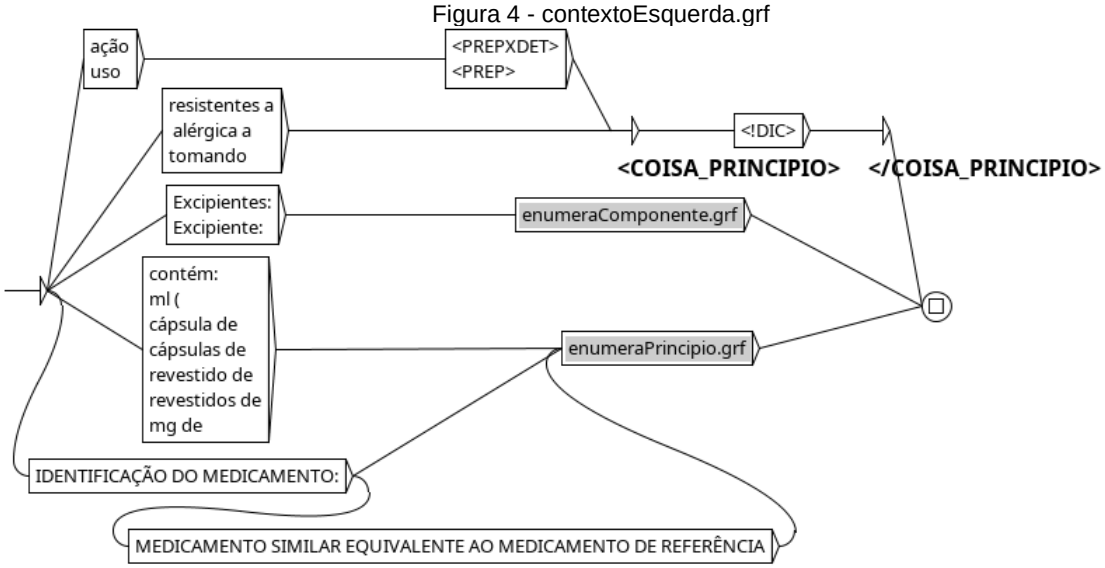
Figura 3 - palavrasDiretas.grf



Fonte: o autor.

A GL *contextoEsquerda.grf*, mostrada na Figura 4, identifica alguns contextos à esquerda observados. Em dois contextos, foi necessário utilizar *tags* lexicais, <*PREP*> que reconhece preposições, e <*PREPXDET*> que indicam preposições com artigos, por exemplo *da* (de + a).

Figura 4 - contextoEsquerda.grf

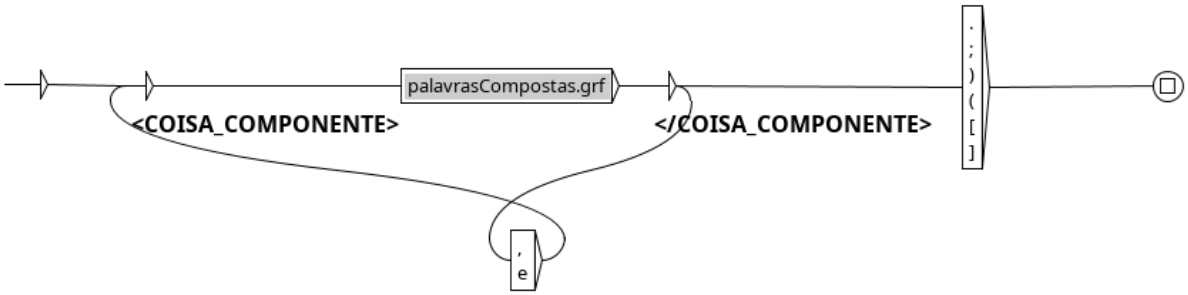


Fonte: o autor.

Foi observado que após alguns contextos a esquerda, existem enumerações, ou seja, indicam uma apresentação sucessiva de substâncias separadas por vírgulas ou conjunções. Então, é importante capturar todas as entidades e diferenciá-las entre componentes e substâncias.

Na Figura 5, é apresentada a GL *enumeraComponente.grf*. Nela há uma subGL *palavrasCompostas.grf,* que verifica se as palavras são compostas, como *citrato de sódio.* As duas GLs responsáveis por identificar as enumerações são semelhantes, diferem-se apenas na anotação de *substância* ou *componente.*

Figura 5 - enumeraComponente.grf



Fonte: o autor.

De forma semelhante a *contextoEsquerda.grf* (Figura 4), foi criada a GL *contextoDireita.grf*, com o objetivo de reconhecer as ENs com base em contextos a direita, que por sua vez é mais difícil de analisar, pois no *corpus* não foram observados padrões tão claros.

A partir dos critérios e métricas já definidos, foram obtidos os resultados das GLs, dispostos na Tabela 2.

Tabela 2 - Resultados

Precisão (%)	Abrangência (%)	Medida-F (%)
74%	32%	45%

Fonte: O autor.

Discussão

No total, foram construídas 6 GLs, com destaque para a *contextoEsquerda.grf* (Figura 4), que se mostrou eficaz por não gerar falso-positivos, ou seja, não reconheceu palavras que não deveriam ser identificadas. Por outro lado, a GL *palavrasDiretas.grf* (Figura 3) apresentou falso-positivos, como *comprimidos revestidos*. Neste exemplo, apenas *comprimidos* foi reconhecido, sem considerar a expressão completa, resultando em um falso-positivo. Portanto, é possível notar que algumas palavras anotadas diretamente podem ser compostas, isto é, precisa-se que ambas palavras sejam reconhecidas para garantir o significado completo.

Sendo assim, deve-se realizar uma validação. Para isso, já existe a GL *palavrasCompostas.grf*, construída para identificar palavras compostas, porém, ela reconhece apenas palavras que não estão no dicionário, o que torna seu uso inviável, visto que na expressão *comprimidos revestidos* ambas palavras estão presentes no dicionário. A melhor abordagem seria anotar diretamente, visto que modificar a GL já existente pode acarretar falso-positivos em outros contextos. Para uma avaliação mais precisa, seria necessário realizar outro estudo linguístico.

De acordo com os números percentuais vistos nos resultados, existe uma margem interessante para melhorar a abrangência. Portanto, há diversas entidades que não estão sendo reconhecidas (falso-negativos), possivelmente porque grande parte não está dentro de um contexto tão claro.

Dito isso, mostrou-se necessário criar mais GLs para abranger melhor o *corpus*, e adaptar as GLs existentes para melhorar a precisão atual. Também é preciso analisar se apenas o uso do *<!DIC>* é suficiente para capturar todas as entidades que estão em algum contexto.

Conclusão

O trabalho demonstra a utilidade do PLN para auxiliar pacientes e médicos a extrair as informações necessárias em bulas. Foi necessário compreender o REN e, posteriormente, realizar um estudo linguístico do *corpus* para observar padrões, que foram necessários para identificar contextos. A metodologia para o REN escolhida foi a linguística. Dessa forma, GLs foram criadas com base nos padrões identificados.

Inicialmente, os resultados mostraram-se satisfatórios. Entretanto, é necessário otimizar a precisão, adaptando as GLs para diminuir os falso-positivos. A métrica da abrangência indica a presença de falso-negativos, palavras que não estão sendo reconhecidas, devido a contextos que não seguem um padrão consistente. O conhecimento de técnicas avançadas com a ferramenta Unitex pode impactar positivamente, especialmente na especificação correta de um padrão.

Para dar sequência ao trabalho, é importante revisar o que já foi feito, melhorar e criar novas GLs. Visto que ainda há algumas palavras que não estão sendo reconhecidas, o ideal é realizar outro estudo linguístico para identificá-las, se possível dentro de um contexto. Por fim, anotá-las e analisar a abrangência adquirida.

Referências

COLOMBO, Cristiano da Silveira; OLIVEIRA, Elias Silva de. Intelligent Information System for Extracting Knowledge from Pharmaceutical Package Inserts. In: **Proceedings of the XVIII Brazilian Symposium on Information Systems**. 2022. p. 1-9.

GROSS, M. **The construction of local grammars**. Finite-state language processing. 1997. p. 329-354.

PIROVANI, J. P. C.. **CRF+ LG: uma abordagem híbrida para o reconhecimento de entidades nomeadas em português**. 2019. Tese (Doutorado em Ciência da Computação) – Centro Tecnológico, Universidade Federal do Espírito Santo, 2019.

PIROVANI, J. P. C.; OLIVEIRA, E. **Estudando a adaptação da NER portuguesa para diferentes gêneros textuais**. *Journal of Supercomputing*, v. 77, p. 13532–13548, 2021. Disponível em: https://doi.org/10.1007/s11227-021-03801-9. Acesso em: 30/06/2024.

SILVA, Tatiane da et al. Bulas de medicamentos e a informação adequada ao paciente. **Revista de Saúde Pública**, v. 34, p. 184-189, 2000.

Agradecimentos

Os autores agradecem à Fundação de Amparo à Pesquisa e Inovação do Espírito Santo (FAPES) pelo financiamento do projeto que deu origem a este trabalho.

Expressamos nossa sincera gratidão à nossa orientadora, Juliana, pela atenção, paciência e orientação, que têm sido essenciais para a realização deste trabalho e para minha formação acadêmica.