GRAMÁTICAS LOCAIS PARA ENTIDADES QUÍMICAS E CONSTRUÇÃO DE *APPLICATION PROGRAMING INTERFACE*

João Henrique Valbusa Lima[1], João Victor Mascarenhas de Faria Santos[2], Juliana Pinheiro Campos Pirovani[2], Elias Silva de Oliveira[1].

[1]Universidade Federal do Espírito Santo, Avenida Fernando Ferrari, 514 - Goiabeiras, 29075-910 - Vitória - ES, Brasil, joao.h.lima@edu.ufes.br, elias@lcad.inf.ufes.br
[2]Universidade Federal do Espírito Santo - UFES (Campus de Alegre) - Alto Universitario, S/N, 29500-000 - Guararema, Alegre - ES, Brasil, joao.vmf.santos@edu.ufes.br, juliana.campos@ufes.br

Resumo
O Reconhecimento de Entidades Nomeadas (REN), é fundamental para atividades que envolvem processamento de texto e extração de informações. Este trabalho tem 2 objetivos: melhorar as Gramáticas Locais (GLs) existentes para o Reconhecimento de Entidades Nomeadas (REN) químicas, tornando-as mais abrangentes; e construir uma *Application Programing Interface* (API) que seja capaz de receber um arquivo de entrada, executar os *scripts* que aplicam tanto as GLs quanto o método híbrido *Conditional Random Fields* (CRF) e GL, e retornar um arquivo com o texto processado e com as entidades nomeadas anotadas. A metodologia inclui um estudo linguístico do domínio químico, estudo das GLs existentes para entidades químicas e alterações nas GLs. A API foi desenvolvida utilizando *Shell Script* e *Makefile* e foi submetida a testes funcionais. Os próximos passos do projeto envolvem a ampliação e refinamento das regras presentes nas GLs, visando maior precisão e abrangência no reconhecimento das entidades químicas, além da disponibilização da API por meio de uma interface *web*.

Palavras-chave: Gramáticas Locais, Reconhecimento de Entidades Nomeadas, Processamento de Linguagem Natural, Entidades Químicas.

Área do Conhecimento: Ciência da Computação.

Introdução

O Processamento de Linguagem Natural (PLN) é um campo da Ciência da Computação, onde sua interação com a Linguística fundamenta o estudo, geração, representação e compreensão da língua natural por computadores. Uma das tarefas centrais dentro do PLN é o Reconhecimento de Entidades Nomeadas (REN), que busca identificar e classificar automaticamente entidades como nomes de pessoas, datas, organizações, locais, dentre outras consideradas relevantes em domínios específicos, como compostos químicos e técnicas de laboratório. Essa tarefa desempenha um papel fundamental em uma variedade de contextos, como na extração de informações em jornais (PIROVANI; OLIVEIRA, 2015), decisões judiciais (VIRTUCIO et al., 2018), patentes (IZO et al., 2023), bulas de remédios (COLOMBO; OLIVEIRA, 2022), entre outros.

As principais abordagens para o REN são baseadas em regras linguísticas e aprendizado de máquina. Na abordagem linguística, são construídas manualmente Gramáticas Locais (GL), compostas por regras que identificam as entidades nomeadas. Na abordagem de aprendizado de máquina, ocorre o treinamento de algoritmos com bases de dados anotadas para reconhecer padrões e classificar as entidades. Pirovani e Oliveira (2015) apresentaram uma abordagem linguística usando GL e Pirovani (2019) apresentou uma abordagem híbrida CRF + GL usando aprendizado de máquina CRF e a abordagem linguística GL para o REN.

Embora já existam GLs para entidades químicas (IZO et al., 2023; IZO et al., 2022) e *scripts* que as aplicam no texto desejado, ainda há necessidade de analisá-las e aprimorá-las, seja por meio de novas regras ou correção de regras existentes. Com a ferramenta Unitex, é possível criar e editar o conjunto de regras de uma GL por meio de grafos.

Além disso, destaca-se que o desenvolvimento de uma API possibilitaria o uso fácil e ágil das ferramentas e *scripts* já existentes que realizam o REN e anotação dessas entidades, com o objetivo de abstrair o conhecimento técnico necessário do código-fonte e chamada dos *scripts*, tornando o seu uso acessível mesmo para aqueles que não são especialistas em computação.

Assim, o objetivo deste trabalho é melhorar as GLs existentes para o reconhecimento de entidades químicas (IZO et al., 2023), além de construir e testar uma API para executar os *scripts* (PIROVANI, 2019) que usam tanto as GLs quanto o método híbrido CRF+GL.

Metodologia

Inicialmente foi realizada uma revisão de literatura de dois trabalhos: a aplicação do método híbrido CRF+GL de Pirovani (2019) e também dos experimentos realizados apenas com GL (PIROVANI; OLIVEIRA, 2015).

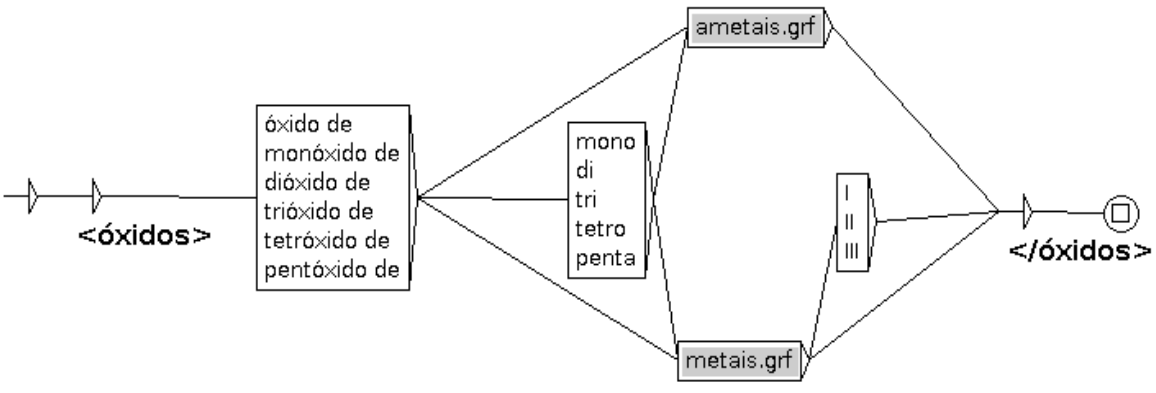
Em seguida, a pesquisa avançou com o estudo dos *scripts* implementados por Pirovani (2019) para aplicação de GLs individualmente e do CRF+GL. Todas as ferramentas necessárias para execução dos *scripts* foram instaladas: Unitex, OpenNLP, Freeling e Mallet. Fez-se uma investigação dos *scripts* que aplicam o método GL, CRF e o híbrido. Os *scripts* recebem um *corpus* de entrada e é gerada uma saída com o texto contendo as entidades nomeadas (EN) anotadas.

Além disso, foi realizado um estudo por meio de seminários e sob a supervisão de especialistas sobre a criação, edição e compilação de GLs através da ferramenta Unitex pela interface gráfica e pela linha de comando.

As regras que compõem uma GL podem ser representadas como grafos. No Unitex, o símbolo do triângulo significa o início e o círculo com um quadrado significa o fim da GL (ver Figura 1). No exemplo da Figura 1 temos uma GL que identifica óxidos. Cada GL pode chamar outras sub-gramáticas como é o caso da Figura 1, que usa as gramáticas que reconhecem metais e ametais. A gramática apresentada possui alguns termos que, se forem reconhecidos no texto, receberão uma marcação indicando a categoria identificada no formato <categoria>termo</categoria>.

Para exemplificar a aplicação da GL, considere a sentença: “O dióxido de nitrogênio e trióxido de enxofre são os principais contribuintes para a formação da chuva ácida”. Após o processamento de texto feito usando a GL de exemplo obtém-se: “O <óxidos>dióxido de nitrogênio </óxidos> e <óxidos>trióxido de enxofre</óxidos> são os principais contribuintes para a formação da chuva ácida”.

Figura 1 - Exemplo de grafo de reconhecimento de óxidos no software Unitex.



Fonte: o Autor

Foi feito um estudo das GLs de Izo (2022) existentes para o contexto de química. Estas gramáticas já realizam o reconhecimento de elementos da tabela periódica, compostos orgânicos, equipamentos, métodos e reações químicas. Ademais, explorou-se regras de formação para certos compostos inorgânicos: óxidos, complexos e sais iônicos e, reações químicas para aplicação em novas regras.

Resultados

Em relação a API, criou-se uma estrutura inicial que tem uma interface que recebe um arquivo de texto de entrada e retorna um arquivo gerado pelos *scripts* com as entidades nomeadas devidamente anotadas.

A primeira versão da API funciona por meio de comandos de terminal do sistema operacional Linux que foram agrupados em um *makefile*. O usuário envia os documentos que serão processados (anotados) no servidor de processamento de PLN com o método SCP. Em seguida, faz-se o acesso remoto ao servidor pelo método SSH e o comando de execução dos *scripts* é chamado. Por fim, o usuário recupera os resultados com o SCP novamente. A Figura 2 mostra o alvo (*target*) "envia" que contém a lista de comandos que executa essas ações no makefile. Vale ressaltar que todo esse processo é abstraído para o usuário com o uso apenas do comando "make".

Figura 2 - Comando executado pelo makefile do usuário.

```
envia:
	sshpass -p $(password) ssh $(userAPI)@$(IP) "cd $(pathAPI); make clean; exit"
	sshpass -p $(password) scp -r $(pathInputUser) $(userAPI)@$(IP):$(inputAPI)
	sshpass -p $(password) ssh $(userAPI)@$(IP) "cd $(pathAPI); make; exit"
	sshpass -p $(password) scp -r $(userAPI)@$(IP):$(outputAPI) $(pathOutputUser)
```

Fonte: o Autor.

O usuário da API tem a opção de fornecer a GL que será usada, ou usar uma gramática padrão da API que reconhece as 10 categorias de entidades nomeadas (EN) do HAREM feita por Pirovani (2019). Além disso, também pode-se escolher usar o método CRF+GL. A Figura 3 mostra o conteúdo de um arquivo de texto em que o usuário queira identificar entidades nomeadas, após o usuário executar o comando da API foi retornado o arquivo da Figura 4. A seguir estão exemplos de como é o funcionamento da API.

Figura 3 - Arquivo enviado pelo usuário para a API

```
{
  "Id": "0123",
  "text": "Meu amigo João me disse que a água é essencial para a vida."
}
```

Fonte: o Autor.

Figura 4 - Arquivo Anotado enviado pela API para o usuário.

```
{
  "Id": "0123",
  "text": "Meu amigo <Pessoa>João</Pessoa> me disse que a
<Composto_Quimico>água</Composto_Quimico> é essencial para a vida."
}
```

Fonte: o Autor.

Neste simples exemplo, a API anotou duas entidades nomeadas: “João” que pertence à categoria pessoa, e “água” que pertence à categoria de composto químico.

Discussão

A API foi testada também por outros alunos que estão adaptando e construindo GLs para outros domínios (decisões judiciais, bulas de medicamentos, expressões cristalizadas, etc.). Verificou-se que a API está funcionando corretamente durante estes testes. Apesar da utilidade e praticidade para

alunos que passaram a utilizar a API, o seu uso ainda está restrito ao nosso grupo de pesquisa. Portanto, trabalhos futuros devem buscar tornar disponível o seu uso em um página na *web*.

As regras que compõem as GLs podem muitas vezes estar suscetíveis a situações identificando os termos de forma inadequada, por isso a adaptação das GLs deve ser feita de forma contínua. Após o estudo das GLs que reconhecem entidades químicas, foi observado que o reconhecimento de métodos e reações químicas podem ser melhorados. Além disso, existe espaço para inclusão de novas regras de formação para certos compostos inorgânicos.

Conclusão

O estudo dos scripts existentes que realizam o REN que permitiu o desenvolvimento da API que irá contribuir para testar GLs em construção e facilitar o uso de GLs pelos usuários. Até agora, o trabalho foi focado no desenvolvimento da API e estudo das GLs existentes para reconhecimento de entidades químicas orgânicas. A próxima etapa do projeto envolverá a criação de GLs específicas para a química inorgânica e aprimoramento das GLs já existentes para métodos, equipamentos e reações químicas. Como trabalho futuro, também pretende-se construir um site para disponibilizar a API para usuários externos.

Referências

APACHE OPENNLP. Disponível em: https://opennlp.apache.org/. Acesso em: 20 set. 2024.

COLOMBO, C. S.; OLIVEIRA, E. Intelligent information system for extracting knowledge from pharmaceutical package inserts. In: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO, 18., 2022, Curitiba. Anais [...]. Curitiba: Sociedade Brasileira de Computação, 2022.

FREELING. Disponível em: https://nlp.lsi.upc.edu/freeling/node/1. Acesso em: 20 set. 2024.

IZO, F.; LEÃO, J.; PIROVANI, J. P. C.; OLIVEIRA, E. Automatic generation of large-scale assessment questions. In: XVIII BRAZILIAN SYMPOSIUM ON INFORMATION SYSTEMS, 2022, Curitiba, Brazil. Anais [...]. New York, NY: Association for Computing Machinery, 2022. Artigo n. 7. DOI: 10.1145/3535511.3535518.

IZO, F.; VEREAU, L. E. S. P.; PIROVANI, J. P. C.; OLIVEIRA, E.; BADUE, C. An intelligent report generator for chemical documents. In: XIX BRAZILIAN SYMPOSIUM ON INFORMATION SYSTEMS, 2023, Maceió, Brazil. Anais [...]. New York, NY: Association for Computing Machinery, 2023. p. 276-283. DOI: 10.1145/3592813.3592915.

MALLET: MACHINE LEARNING FOR LANGUAGE TOOLKIT. Disponível em: https://mallet.cs.umass.edu/download.php. Acesso em: 20 set. 2024.

PIROVANI, J. P. C. CRF+LG: uma abordagem híbrida para o reconhecimento de entidades nomeadas em português. 2019. 212 f. Tese (Doutorado em Informática) – Universidade Federal do Espírito Santo, Vitória, ES, 2019.

PIROVANI, J. P. C.; NOGUEIRA, M.; OLIVEIRA, E. Indexing names of persons in a newspaper large dataset. In: PROPOR, 13., 2018, Canela, RS. Anais [...]. Springer, 2018. p. 11122. DOI: 10.1007/978-3-319-99722-3_15.

PIROVANI, J. P. C.; OLIVEIRA, E. Extração de nomes de pessoas em textos em português: uma abordagem usando gramáticas locais. In: COMPUTER ON THE BEACH 2015, Florianópolis, SC. Anais [...]. Florianópolis, SC: SBC, 2015.

PIROVANI, J. P. C.; OLIVEIRA, E. Studying the adaptation of Portuguese NER for different textual genres. Journal of Supercomputing, v. 77, p. 13532-13548, 2021.

UNITEX/GRMLAB. Unitex-GramLab 3.3 User Manual. Disponível em: https://unitexgramlab.org/releases/3.3/man/Unitex-GramLab-3.3-usermanual-en.pdf. Acesso em: 20 set. 2024.

VIRTUCIO, Michael Benedict L. et al. Predicting Decisions of the Philippine Supreme Court Using Natural Language Processing and Machine Learning. In: 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), vol. 02, p. 130-135, 2018. DOI: 10.1109/COMPSAC.2018.10348.

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.