UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI

Julia Luiza Ferreira Santos

Criação de Grafos de Conhecimento Baseados na Extração de Entidades e Relações em Textos de Imunologia

São João del-Rei

2025

UNIVERSIDADE FEDERAL DE SÃO JOÃO DEL-REI

Julia Luiza Ferreira Santos

Criação de Grafos de Conhecimento Baseados na Extração de Entidades e Relações em Textos de Imunologia

Monografia apresentada como requisito da disciplina de Projeto Orientado em Computação I do Curso de Bacharelado em Ciência da Computação da UFSJ.

Orientador: Alexandre Bittencourt Pigozzo

Universidade Federal de São João del-Rei – UFSJ

Bacharelado em Ciência da Computação

São João del-Rei

2025

Lista de abreviaturas e siglas

PLN	Processamento de Linguagem Natural
LLM	Modelos de Linguagem em Grande Escala (Large Language Model)
RAG	Geração Aumentada de Recuperação (Retrieval-Augmented Generation)
API	Interface de Programação de Aplicações
NER	Reconhecimento de Entidades Nomeadas (NER)
RE	Extração de Relações (RE)

Sumário

1 Introdução

Os grafos de conhecimento possuem inúmeras aplicações, como sistemas de recomendação, detecção de fake news e sistemas de diálogo (1). Um grafo é representado na forma de triplas (Entidade1 – Relação – Entidade2), onde os nós representam entidades ou conceitos e as arestas representam os relacionamentos entre eles. Ele exibe o conhecimento estruturado por meio de uma interface gráfica, permitindo que os usuários visualizem e descubram, de forma mais intuitiva, as relações entre as informações (2). Aplicações bem-sucedidas de grafos de conhecimento já podem ser observadas na área médica. Grafos médicos integram dados provenientes de múltiplas fontes e auxiliam em tarefas como busca de informações, diagnóstico, tratamento e prognóstico (3). Por analogia, um grafo voltado à imunologia pode não apenas estruturar o conhecimento básico da área, mas também apoiar análises sobre interações imunológicas e interpretação de processos complexos.

A construção de grafos de conhecimento, porém, enfrenta algumas limitações. Em áreas altamente especializadas, pode ser necessária a participação de um grande número de especialistas para definir corretamente as entidades e relações relevantes. Ainda que os avanços em Processamento de Linguagem Natural (PLN) tenham reduzido essa dependência, grafos construídos a partir de ferramentas de PLN geralmente exigem grandes quantidades de dados anotados para treinar modelos com boa precisão. Nesse sentido, os LLMs contribuem significativamente para superar esse desafio, uma vez que permitem extrair entidades nomeadas e relações sem demandar grande esforço humano ou dados adicionais de treinamento. Isso ocorre porque técnicas como o fine-tuning podem aprimorar o desempenho dos modelos em tarefas textuais (4). Estudos recentes têm demonstrado que a adição de apenas uma única demonstração de tarefa ao prompt aumenta de forma expressiva a qualidade da extração de triplas, e que o método Geração Aumentada de Recuperação (RAG - *Retrieval Augmented Generation*) contribui para elevar a precisão dessas extrações (5).

Com base nessas observações, este trabalho propõe um sistema para a criação de grafos de conhecimento a partir de artigos científicos da área da imunologia. O objetivo é estruturar o conhecimento presente nesses materiais de modo a evidenciar relações e interações biológicas entre entidades imunológicas. Para isso, utilizou-se a API gemini-2.5-flash para extrair quadriplas de conhecimento — compostas por entidade1, relação, entidade2 e uma condição que especifica quando a relação ocorre. A extração é guiada por um prompt que fornece ao modelo alguns exemplos de textos e suas quádruplas. Esses exemplos são recuperados pelo método RAG a partir de uma pequena base de dados construída com anotações relizadas manualmente.

2 Referencial Teórico

Esta seção apresenta conceitos importantes para a compreensão do trabalho desenvolvido.

2.1 Modelos de Linguagem de Grande Escala (LLMs)

Modelos de linguagem de grande escala (LLMs, do inglês *Large Language Models*) são sistemas de inteligência artificial (IA) treinados com uma grande quantidade de dados de alta qualidade. Temos como principais exemplos o GPT-3 e GPT-4, os modelos Llama da Meta e o Gemini da Google. Esses modelos utilizam arquitetura de redes neurais e aprendizado profundo (*deep learning*) para aprender como as palavras são usadas em conjunto na linguagem. Os padrões aprendidos são, então, aplicados para representar as complexas relações entre palavras (6). Dessa forma, os LLMs se tornaram capazes de gerar textos de forma semelhante à linguagem humana, interpretando contextos, gerando respostas coerentes, e realizando tarefas de tradução e resumo (7). Essa área do conhecimento é chamada de processamento de linguagem natural (PLN).

Os modelos de linguagem de grande escala são baseados em "*transformers*", sendo sua principal característica a presença de mecanismos de autoatenção(8). Um mecanismo de atenção determina a importância relativa de diferentes partes da sequência de entrada, influenciando o modelo a considerar o que é importante e desconsiderar o que não é. Matematicamente, os modelos calculam "pesos"para aumentar ou diminuir a influência de cada parte de uma sequência de entrada, de acordo com a sua importância(9).

Os avanços no poder computacional, na disponibilidade de conjuntos de treinamento robustos e no desenvolvimento de técnicas de aprendizado profundo contribuíram diretamente para o surgimento de LLMs que são capazes de reconhecer, interpretar e gerar textos com nenhum (*zero-shot*) ou poucos (*few-shots*) exemplos de realização da tarefa dada ao modelo (6). Por serem capazes de lidar eficientemente com uma ampla variedade de tarefas, os LLMs eliminam a necessidade de desenvolver e treinar modelos específicos para cada domínio ou área do conhecimento, o que seria um processo limitado por custos e restrições de recursos (10).

2.2 Grafos de Conhecimento

2.3 Extração de Relacionamentos

A extração de relacionamentos (RE) é uma tarefa no processamento de linguagem natural (NLP) que envolve extrair entidades de um texto e as relações entre elas. Em relações binárias, por exemplo, cada relação é representada como triplas: (entidade1, relacionamento, entidade2) (11). Dessa maneira, poderíamos, por exemplo, representar a habilidade do sistema imunológico de defender o organismo contra o SARS-CoV-2, pela tripla: *immunologic system*, *defends-against*, SARS-CoV-2.

RE tem o potencial de estruturar dados não estruturados a fim de gerar um conhecimento relevante. A extração de relacionamentos pode ser feita a nível de sentença ou a nível de documento. Neste último caso, a extração de entidades e suas relações é feita através de uma análise cruzada de sentenças (11, 12), refletindo melhor a realidade, uma vez que, em cenários comuns, fatos relacionados são expressos entre múltiplas sentenças (13, 14).

Identificar vínculos entre entidades de um texto pode ser feito usando técnicas baseadas em regras, aprendizado de máquina tradicional (*Machine Learning* - ML) e aprendizado profundo (15). Métodos baseados em regras são interpretáveis, mas exigem muito esforço manual e têm baixa generalização, o que limita seu uso em grandes volumes de dados (16). Uma vez que os métodos de ML tradicionais exigem conhecimento especializado do domínio para definir características, a extração de relações passa a ser uma tarefa que melhor adapta-se a técnicas de aprendizado profundo, visto que redes neurais (NN) extraem características de dados brutos, sem a necessidade de conhecimento especializado, superando as demais abordagens no que diz respeito à generalização para vários conjuntos de dados e domínios (15).

O Reconhecimento de Entidades Nomeadas (NER) e a Extração de Relações (RE) têm apresentado melhorias significativas com os avanços em Processamento de Linguagem Natural (PLN), especialmente em arquiteturas baseadas em *transformers*. Modelos como o BERT e o GPT, que utilizam *embeddings* contextuais de palavras, são exemplos desses avanços (17).

Existem duas abordagens principais para a extração de relações: a RE baseada em pipeline e a extração conjunta. A primeira divide o processo em etapas: primeiro, são identificadas as entidades presentes no texto e, posteriormente, são detectadas as relações entre elas. Dessa forma, o modelo estabelece relações apenas entre pares de entidades previamente reconhecidos. No entanto, essa abordagem enfrenta o risco de propagação de erros, caso equívocos sejam introduzidos na etapa de identificação de entidades. A segunda abordagem, como o nome indica, realiza a identificação de entidades e classificação de

relacionamentos de forma conjunta. A propagação de erros nessa abordagem é minimizada porque, dentre outras razões, ao aprender conjuntamente as tarefas, os modelos conjuntos podem se adaptar a erros em uma tarefa aproveitando informações da outra. (11).

2.4 Engenharia de Prompt

Prompts servem como instruções para os LLMs, direcionando-os a gerar as saídas desejadas. A sua qualidade interfere diretamente na precisão das respostas geradas pelo modelo, portanto a Engenharia de *Prompt* é fundamental para se obter resultados satisfatórios dos modelos de linguagem.

In-context learning (ICL) é um método no qual o modelo "aprende por demonstração". A partir de exemplos fornecidos, o modelo aprende a reconhecer e replicar tarefas. Ele ajusta suas respostas para se alinharem aos exemplos demonstrados, eliminando a necessidade de treinamento e ajuste no modelo. Dessa forma, ao receberem uma sequência de entrada (um *prompt* textual, por exemplo), os LLMs podem aprender a executar tarefas de extração de conhecimento (5) e, no caso deste trabalho, de relações entre entidades.

O desempenho do modelo tende a ser melhor quando o *prompt* é combinado com a descrição da tarefa e um exemplo que demonstra a execução da tarefa descrita. Um *prompt* que requer que o modelo execute uma determinada tarefa, sem nenhuma demonstração desta, é chamado *prompts zero-shot* (Figura 1). Neste caso, o modelo usa o seu conhecimento interno na execução da tarefa. Quando é fornecido um único exemplo ao modelo, juntamente com a instrução, temos um *prompt one-shot* (Figura 2). A partir de três exemplos, trata-se de um *prompt few-shot*(5).

```
Extract knowledge triples from the text. Return the triples in JSON format.
Text: The Porsche Panamera is a mid/full-sized luxury vehicle (E-segment in Europe)
manufactured by the German automobile manufacturer Porsche. It is front-engined and
has a rear-wheel-drive layout, with all-wheel drive versions also available.
Your answer:
```

Figura 1 – Exemplo de prompt zero-shot extraído do artigo de Polat et al.(2024) (5).

```
Extract knowledge triples from the text. Return the triples in JSON format.
Here is an example.
Text: The Amazon River flows through Brazil and Peru.
Your answer: {"Triples": [["Amazon River", "country", "Brazil"], ["Amazon River",
"country", "Peru"]]}
Text: The Porsche Panamera is a mid/full-sized luxury vehicle (E-segment in Europe)
manufactured by the German automobile manufacturer Porsche. It is front-engined and
has a rear-wheel-drive layout, with all-wheel drive versions also available.
Your answer:
```

Figura 2 – Exemplo de prompt one-shot extraído do artigo de Polat et al.(2024) (5).

2.5 Recuperação Aumentada por Geração (RAG)

Os modelos de linguagem neural pré-treinados aprendem a partir de um conjunto de dados, sem acesso à memória externa, isto é, acessam apenas sua memória paramétrica. Esses modelos, no entanto, apresentam desvantagens, como não expandir ou revisar seu conhecimento facilmente. Uma forma de enfrentar essa limitação é pelo uso da memória paramétrica em combinação com mecanismos de recuperação de informações em uma base de conhecimento (memória não paramétrica) (18). Esse mecanismo é chamado Recuperação Aumentada por Geração (RAG).

RAG é, portanto, uma técnica de ajuste fino usada em LLMs que combina memórias paramétrica e não paramétrica. Sua estrutura básica consiste em: uma base de conhecimento, um codificador de *embeddings*, um mecanismo de recuperação e um gerador. A base de conhecimento é incorporada pelo codificador de *embeddings*, e o mecanismo de recuperação realiza uma busca nessa base, recuperando dela a informação/contexto cujo *embedding* possui maior semelhança com o *embedding* do texto fornecido ao modelo(5).

Em *prompts one-shot* e *few-shot*, em vez de manter exemplos fixos, ou selecioná-los aleatoriamente, o RAG pode ser aplicado para incluir no prompt os exemplos de resolução da tarefa que mais se assemelham à tarefa em questão(5). Essa é uma variação do modelo *in-context learning*, em que o modelo aprende a partir de exemplos no próprio prompt.

2.6 Limitação dos métodos de avaliação

Conforme Polat et al.(2024)(5), métodos rígidos de avaliação da extração de relações se adaptam melhor a modelos menores. Modelos como o BART, treinados em um grande conjunto de dados e submetidos a um extenso processo de ajuste fino, conseguem produzir saídas padronizadas de forma consistente. No entanto, a expressividade dos LLMs e sua capacidade de variabilidade no uso de expressões dificultam a tarefa de avaliação das relações extraídas. Eles são capazes de produzir diversos tipos de saída com conteúdo semelhante ao conteúdo alvo, uma vez que não simplesmente classificam tokens de entrada com base em um conjunto predefinido de classes, mas geram novos tokes a partir de um vocabulário extenso. Dessa forma, considerando a natureza aberta das saídas dos LLMs, profissionais frequentemente optam pela avaliação humana, embora seja uma abordagem mais demorada e mais cara.

3 Trabalhos Relacionados

Alguns trabalhos serviram de inspiração ou referência para o desenvolvimento da metodologia deste Trabalho de Conclusão de Curso (TCC). O primeiro a ser destacado é o TCC do Rodrigues (2025) (19), que aborda o uso de modelos generativos (LLaMA 3.2:3B e o DeepSeek-R1:8B) para extração de entidades nomeadas de textos da área de Imunologia e de textos sobre modelagem computacional do sistema imune.

Rodrigues (2025) propôs um pipeline, cuja primeira etapa consiste na segmentação de um arquivo PDF de entrada para obtenção de trechos usados como tarefas para o modelo. A tarefa é expressa em um prompt construído de modo a restringir o uso de termos a uma lista pré-definida de entidades nomeadas. Para melhorar o desempenho do modelo, o sistema inclui no prompt uma contextualização do assunto tratado no texto de entrada do LLM. Para isso, é utilizado o método RAG, que recupera do banco de dados o contexto cuja incorporação vetorial mais se assemelha à do texto da tarefa. O banco de dados escolhido foi o MongoDB, devido à sua capacidade de busca vetorial otimizada (Vector Search). Esse trabalho, no entanto, não empreende a tarefa de extração de relações entre as entidades obtidas. Além disso, não explora a capacidade dos LLMs generativos de criar novos tokens — recurso que poderia ser vantajoso em casos em que as entidades nomeadas e as relações são apresentadas de forma implícita no texto.

O trabalho de Polat et al. (2024)(5), em contrapartida, testa diferentes métodos de engenharia de prompt nos modelos GPT-4, Mistral 7B e Llama 3. Em vez de estabelecer previamente um escopo fixo de possíveis relações, os autores permitem que o modelo extraia relações para além de correspondências exatas, o que resulta em inferências mais refinadas a partir dos textos. Nesse trabalho, o RAG foi utilizado para selecionar exemplos de tarefas semelhantes à tarefa dada ao modelo.

Outro diferencial do artigo foi a sua proposta de um método de avaliação baseado em ontologias. O objetivo é reduzir a dependência da avaliação humana, aproveitando a semântica dos dados e as restrições de propriedades presentes no Wikidata(20) para automatizar o processo de validação das triplas extraídas. No entanto, a ontologia do Wikidata é dinâmica e edições realizadas por usuários podem afetar a avaliação. O conteúdo existente na base também é propenso a erros, não podendo ser assumido como verdade absoluta. Além disso, a técnica utilizada no trabalho para vincular uma entidade obtida a um termo existente no Wikidata também apresentou limitações, uma vez que algumas correspondências geradas não foram corretas(5). Apesar dessas limitações, a proposta de avaliação mostrou-se promissora e com potencial de aperfeiçoamento.

Já o trabalho de Xu et al. (2024) (3) utiliza o modelo ChatGPT 3.5-Turbo-16k API

nas tarefas de Reconhecimento de Entidades Nomeadas (NER) e Extração de Relações (RE). Diferentemente dos estudos anteriormente citados, o método proposto também realiza a desambiguação de entidades, removendo ambiguidades e sinônimos por meio da métrica de similaridade de Jaccard. A partir das entidades e relações extraídas, os autores constroem um grafo de conhecimento sobre insuficiência cardíaca, o qual pode auxiliar na tomada de decisões clínicas e facilitar a visualização de informações relevantes para diagnóstico, tratamento e prognóstico.

De modo geral, os estudos analisados trazem contribuições importantes para o uso de modelos de linguagem de grande porte em tarefas de extração de informação, especialmente por meio de suas metodologias de engenharia de prompts e do uso do RAG. Entre eles, o trabalho de Rodrigues (2025) destaca-se por aplicar modelos generativos à extração de entidades nomeadas em textos biomédicos, embora não aborde a extração de relações. Essa limitação marca o primeiro ponto de partida para o desenvolvimento deste TCC. Além disso, nota-se que nenhum dos estudos explora a identificação das condições em que as relações ocorrem, o que configura o segundo aspecto inovador desta proposta. Ambos os pontos são detalhados na seção de Metodologia.

4 Metodologia

Com o objetivo de estruturar o conhecimento presente em textos da área de imunologia, esta metodologia utiliza o Reconhecimento de Entidades Nomeadas (NER) e a Extração de Relações (RE), apoiadas em técnicas de engenharia de prompt aplicadas a um modelo de linguagem de grande escala (LMM), inspirado em Polat et al. (2024). Para a obtenção de exemplos relevantes ao modelo, empregou-se o método de Geração Aumentada de Recuperação (RAG), utilizando a pesquisa vetorial do MongoDB como mecanismo de busca semântica. Além disso, com o intuito de facilitar a compreensão das interações biológicas, o sistema desenvolvido gera um grafo de conhecimento que representa visualmente as relações identificadas entre as entidades.

A metodologia foi organizada em cinco etapas principais: (1) preparação da base de dados; (2) leitura e segmentação de artigos; (3) aplicação do RAG; (4) construção do prompt; e (5) geração do grafo de conhecimento. Cada uma dessas etapas será detalhada nas subseções a seguir.

4.1 Base de dados

O banco de dados utilizado neste TCC foi o MongoDB. A escolha se deve à disponibilidade do recurso de pesquisa vetorial (*Vector Search*). Diferentemente da busca tradicional, que localiza correspondências exatas de texto, a pesquisa vetorial identifica vetores próximos à consulta em um espaço multidimensional (21). Para isso, as incorporações vetoriais do texto de entrada são comparadas com as demais armazenadas no banco de dados. Diferentes tipos de métricas podem ser utilizadas para o cálculo da proximidade vetorial entre as incorporações de texto. Neste trabalho, utilizou-se a similaridade do cosseno.

A base de dados é utilizada para armazenar exemplos de tarefas semelhantes às que o modelo deve executar. Assim, cada registro contém um trecho de um artigo da área de imunologia, sua incorporação vetorial (*embedding*) e uma lista de quádruplas. Cada quádrupla contém entidades nomeadas identificadas no texto, as relações entre elas e as condições em que essas relações ocorrem. O banco de dados foi anotado manualmente primeiro foram extraídos trechos de textos da área da imunologia (22, 23) e, posteriormente, os trechos foram análisados para a obtenção das relações existentes em cada um deles.

As incorporações foram geradas utilizando a ferramenta *Sentence Transformers* (SBERT), um módulo *Python* que possibilita acessar, usar e treinar modelos de embed-

ding. Essa ferramenta permite calcular representações vetoriais densas para sentenças e parágrafos (24). O modelo adotado foi o *all-MiniLM-L6-v2*, que mapeia textos para um espaço vetorial de 384 dimensões, sendo adequado para tarefas como agrupamento e busca semântica (25).

A seguir, apresenta-se parte de um texto da área de imunologia(26), bem como suas quádruplas anotadas manualmente para compor a base de dados de exemplos para o modelo:

"The CD8-TL recognizes intracytoplasmic antigens presented by MHC molecules class I, which are expressed by practically all nucleated cells. ... At a subsequent contact, the CD8-TL can eliminate by cytotoxicity any cell that presents this specific antigen. The CD8-TL induce apoptosis in the target cell through the action of perforins and granzymes and can also lead to apoptosis through the expression of the Fas L receptor. ... The TL ... can recognize antigens even in the absence of presentation by the MHC molecule."


```
[
  {
    "source": "CD8-TL",
    "relation": "recognizes",
    "destiny": "intracytoplasmic antigens",
    "condition": "presented by MHC class I molecules"
  },
  {
    "source": "MHC class I molecules",
    "relation": "expressed_by",
    "destiny": "nucleated cells",
    "condition": ""
  },
  {
    "source": " TL",
    "relation": "recognizes",
    "destiny": "antigens",
    "condition": "even in the absence of presentation by the MHC molecule"
  },
  {
    "source": "CD8-TL",
    "relation": "eliminates",
    "destiny": "target cell",
    "condition": "presence of specific antigen and previous contact"
  },
```

```
  {
    "source": "CD8-TL",
    "relation": "induces_apoptosis_in",
    "destiny": "target cell",
    "condition": "through perforins and granzymes action OR
    Fas L receptor interaction"
  }
]
```

4.2 Segmentação de artigos

Muitos estudos se concentram em extrair relações a partir de sentenças individuais. Entretanto, muitas aplicações do mundo real exigem que os sistemas de RE identifiquem entidades e relações que ultrapassem os limites das frases. A linha de pesquisa em RE voltada à extração de relações em nível de documento apresenta a vantagem de capturar de forma mais completa as informações relacionais distribuídas ao longo de um texto (11).

Considerando a relevância das análises que envolvem múltiplas sentenças — e a capacidade dos LLMs de lidar com esse tipo de processamento — neste trabalho, o processo de RE e NER foi conduzido em nível de parágrafo, e cada um dos trechos foi utilizado para a criação de um grafo de conhecimento de todo o documento. Para isso, utilizou-se a biblioteca *pdfplumber* para a leitura do conteúdo dos artigos e sua posterior segmentação em parágrafos. A escolha de trabalhar com parágrafos foi motivada pela necessidade de preservar o contexto e as relações implícitas entre as entidades.

4.3 RAG

O método RAG consiste em fornecer ao modelo exemplos relevantes selecionados a partir da similaridade semântica com a tarefa dada. Sua estrutura é composta por quatro elementos principais: uma base de conhecimento, um codificador de embeddings, um mecanismo de recuperação e um gerador (5). A base de conhecimento e o processo de codificação dos textos em embeddings — utilizados para identificar o exemplo mais semelhante por meio da busca vetorial — foram apresentados na Seção 4.1.

De forma resumida, o trecho extraído do artigo na etapa de segmentação é primeiro convertido em *embeddings*. Em seguida, o mecanismo de busca vetorial localiza, na base de dados, os três exemplos mais similares ao trecho em questão a comparação é feita entre a incorporação vetorial do trecho extraído do artigo e a incorporação de cada um dos textos da base de dados. Esses exemplos são, então, adicionados ao prompt, que descreve a tarefa a ser realizada pelo modelo. A etapa de engenharia de prompts é detalhada

na subseção seguinte. A partir do prompt, o modelo interpreta as instruções e realiza a análise solicitada, extraindo as entidades nomeadas presentes no trecho, identificando suas relações e determinando as condições necessárias para que essas relações ocorram. Para esse processo, foi utilizado o modelo *gemini-2.5-flash*.

4.4 Prompt

A técnica de Engenharia de Prompt utilizada foi a *few-shot prompting*, que consiste em apresentar ao modelo alguns exemplos da tarefa a ser executada no caso deste trabalho, três exemplos. Um exemplo completo de prompt está disponível na seção de Apêndices deste trabalho. Ele foi estruturado para cumprir as seguintes funções:

- definir o papel do modelo;
- indicar os elementos que devem ser ignorados no texto;
- contextualizar a área de atuação (imunologia);
- descrever claramente a tarefa (extração de quádruplas);
- fornecer exemplos da tarefa (obtidos via método RAG);
- apresentar o trecho de texto sobre o qual o modelo deve atuar.

O papel atribuído ao modelo é o de **analisador especializado em extrair relações biológicas**. Seu objetivo é identificar apenas interações biologicamente relevantes, representadas como relações entre células, moléculas ou processos imunológicos. Para evitar ruídos, o modelo foi instruído a ignorar relações que não pertencem ao domínio biológico, como relações sociais, históricas ou culturais.

A saída esperada foi definida no formato `JSON`, com cada quádrupla composta pelos seguintes campos:

- **source**: entidade de origem da relação;
- **relation**: verbo que expressa a interação;
- **destiny**: entidade de destino da relação;
- **condition**: condição necessária para que a relação ocorra, inferida pelo modelo a partir do contexto.

4.5 Grafo de Conhecimento e Interface Gráfica

Para a geração do grafo de conhecimento a partir dos `JSONs` produzidos pelo modelo, foi utilizada a biblioteca *Pyvis*, responsável por criar grafos interativos em formato `HTML`. Já para o desenvolvimento da interface gráfica com o usuário, optou-se pela biblioteca *Streamlit*.

5 Resultados

6 Conclusão

Referências

1 QUIROZ, R. N.; CAMACHO, J. V.; PEñATA, E. Z.; LEMUS, Y. B.; LóPEZ-FERNáNDEZ, C.; ESCORCIA, L. G.; FERNáNDEZ-PONCE, C.; COBOS, M. R.; MORENO, J. F.; FIORILLO-MORENO, O.; QUIROZ, E. N. Multiscale information processing in the immune system. *Frontiers in Immunology*, v. 16, 2025. Disponível em: <https://www.frontiersin.org/journals/immunology/articles/10.3389/fimmu.2025.1563992>. Citado na página 4.

2 HAO, X.; JI, Z.; LI, X.; YIN, L.; LIU, L.; SUN, M.; LIU, Q.; YANG, R. Construction and application of a knowledge graph. *Remote Sensing*, v. 13, n. 13, p. 2511, 2021. Disponível em: <https://doi.org/10.3390/rs13132511>. Citado na página 4.

3 XU, T.; GU, Y.; XUE, M.; GU, R.; LI, B.; GU, X. Knowledge graph construction for heart failure using large language models with prompt engineering. *Frontiers in Computational Neuroscience*, v. 18, p. 1389475, 2024. Citado 2 vezes nas páginas 4 e 9.

4 YE, Y. Construction and application of materials knowledge graph in multidisciplinary materials science via large language model. *Advances in Neural Information Processing Systems*, v. 37, 2024. Citado na página 4.

5 POLAT, F.; TIDDI, I.; GROTH, P. Testing prompt engineering methods for knowledge extraction from text. *Semantic Web*, 2024. Disponível em: <https://www.researchgate.net/publication/384205838_Testing_prompt_engineering_methods_for_knowledge_extraction_from_text>. Citado 5 vezes nas páginas 4, 7, 8, 9 e 13.

6 THIRUNAVUKARASU, A. J.; TING, D. S. J.; ELANGOVAN, K.; GUTIERREZ, L.; TAN, T. F.; TING, D. S. W. Large language models in medicine. *Nature Medicine*, v. 29, p. 1930–1940, 2023. Citado na página 5.

7 LIU, Y.; HAN, T.; MA, S.; ZHANG, J.; YANG, Y.; TIAN, J.; HE, H.; LI, A.; HE, M.; LIU, Z.; WU, Z.; ZHAO, L.; ZHU, D.; LI, X.; QIANG, N.; SHEN, D.; LIU, T.; GE, B. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, v. 1, n. 2, p. 100017, 2023. ISSN 2950-1628. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2950162823000176>. Citado na página 5.

8 STRYKER, C.; BERGMANN, D. *What is a transformer model?* 2025. Accessed: 2025-10-06. Disponível em: <https://www.ibm.com/think/topics/transformer-model>. Citado na página 5.

9 BERGMANN, D.; STRYKER, C. *What is an attention mechanism?* 2025. Accessed: 2025-10-06. Disponível em: <https://www.ibm.com/think/topics/attention-mechanism>. Citado na página 5.

10 RUAN, W.; LYU, Y.; ZHANG, J.; CAI, J.; SHU, P.; GE, Y.; LU, Y.; GAO, S.; WANG, Y.; WANG, P.; ZHAO, L.; WANG, T.; LIU, Y.; FANG, L.; LIU, Z.; LIU, Z.; LI, Y.; WU, Z.; CHEN, J.; JIANG, H.; PAN, Y.; YANG, Z.; CHEN, J.; LIANG, S.;

ZHANG, W.; MA, T.; DOU, Y.; ZHANG, J.; GONG, X.; GAN, Q.; ZOU, Y.; CHEN, Z.; QIAN, Y.; YU, S.; LU, J.; SONG, K.; WANG, X.; SIKORA, A.; LI, G.; LI, X.; LI, Q.; WANG, Y.; ZHANG, L.; ABATE, Y.; HE, L.; ZHONG, W.; LIU, R.; HUANG, C.; LIU, W.; SHEN, Y.; MA, P.; ZHU, H.; YAN, Y.; ZHU, D.; LIU, T. *Large Language Models for Bioinformatics*. 2025. Disponível em: <https://arxiv.org/abs/2501.06271>. Citado na página 5.

11 ZHAO, X.; DENG, Y.; YANG, M.; WANG, L.; ZHANG, R.; CHENG, H.; LAM, W.; SHEN, Y.; XU, R. A comprehensive survey on relation extraction: Recent advances and new frontiers. *ACM Computing Surveys*, v. 56, n. 11, p. 293, 2024. Disponível em: <https://doi.org/10.1145/3674501>. Citado 3 vezes nas páginas 6, 7 e 13.

12 XU, T.; QU, J.; HUA, W.; LI, Z.; XU, J.; LIU, A.; ZHAO, L.; ZHOU, X. Evidence reasoning and curriculum learning for document-level relation extraction. *IEEE Transactions on Knowledge and Data Engineering*, v. 36, n. 2, p. 594–607, 2024. Citado na página 6.

13 SHANG, Y.; GUO, Y.; HAO, S.; HONG, R. *Biomedical Relation Extraction via Adaptive Document-Relation Cross-Mapping and Concept Unique Identifier*. 2025. Disponível em: <https://arxiv.org/abs/2501.05155>. Citado na página 6.

14 YAO, Y.; YE, D.; LI, P.; HAN, X.; LIN, Y.; LIU, Z.; LIU, Z.; HUANG, L.; ZHOU, J.; SUN, M. DocRED: A large-scale document-level relation extraction dataset. In: KORHONEN, A.; TRAUM, D.; MÀRQUEZ, L. (Ed.). *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, 2019. p. 764–777. Disponível em: <https://aclanthology.org/P19-1074/>. Citado na página 6.

15 MOUICHE, I.; MERBOUH, H.; SAAD, S. *Context-aware Entity-Relation Extraction Pipeline for Threat Intelligence Knowledge Graphs*. [S.l.], jan. 2025. Disponível em: <https://doi.org/10.36227/techrxiv.173627493.39916970/v1>. Citado na página 6.

16 LIU, Z.; CHEN, X.; WANG, H.; LIU, X. Integrating regular expressions into neural networks for relation extraction. *Expert Systems with Applications*, v. 252, p. 124252, 2024. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417424011187>. Citado na página 6.

17 FABACHER, T.; SAULEAU, E.-A.; ARCAY, E.; FAYE, B.; ALTER, M.; CHAHARD, A.; MIRAILLET, N.; COULET, A.; NéVéOL, A. *Efficient extraction of medication information from clinical notes: an evaluation in two languages*. 2025. Disponível em: <https://arxiv.org/abs/2502.03257>. Citado na página 6.

18 LEWIS, P.; PEREZ, E.; PIKTUS, A.; PETRONI, F.; KARPUKHIN, V.; GOYAL, N.; KüTTLER, H.; LEWIS, M.; YIH, W.-t.; ROCKTäSCHEL, T.; RIEDEL, S.; KIELA, D. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2005.11401*, 2020. Citado na página 8.

19 RODRIGUES, G. M. *Desenvolvimento de um prompt para o reconhecimento de entidades nomeadas usando modelos de inteligência artificial generativa*. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) — Universidade Federal de São João del-Rei (UFSJ), São João del-Rei, 2025. Citado na página 9.

20 VRANDECIC, D.; KRÖTZSCH, M. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, v. 57, n. 10, p. 78–85, 2014. Citado na página 9.

21 MONGODB. *MongoDB Vector Search Overview*. 2025. <https://www.mongodb.com/docs/atlas/atlas-vector-search/vector-search-overview/>. Accessed: 2025-02-17. Citado na página 11.

22 SOMPAYRAC, L. *How the Immune System Works*. 6. ed. [S.l.]: Wiley Blackwell, 2019. 145 p. Soft cover; includes index. Citado na página 11.

23 JúNIOR, D. M.; ARAúJO, J. A.; CATELAN, T. T.; SOUZA, A. W.; CRUVINEL, W. M.; ANDRADE, L. E.; SILVA, N. P. Immune system - part ii: basis of the immunological response mediated by t and b lymphocytes. *Revista Brasileira de Reumatologia*, v. 50, n. 5, p. 552–580, Sep-Oct 2010. Citado na página 11.

24 SENTENCETRANSFORMERS Documentation. 2025. Created with Sphinx using a theme provided by Read the Docs. Disponível em: <https://sbert.net/>. Citado na página 12.

25 SENTENCE-TRANSFORMERS. *all-MiniLM-L6-v2*. 2021. <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>. Accessed: 2025-02-17. Citado na página 12.

26 DE, S.; SANYAL, D. K.; MUKHERJEE, I. Fine-tuned encoder models with data augmentation beat chatgpt in agricultural named entity recognition and relation extraction. *Expert Systems with Applications*, v. 277, p. 127126, 2025. ISSN 0957-4174. Disponível em: <https://doi.org/10.1016/j.eswa.2025.127126>. Citado na página 12.

Apêndices

APÊNDICE A – Prompt

You are an analyzer and your role is to perform Relation Extraction. Your objective is to extract meaningful relationships from biological texts. The relationships represent biological interactions among cells or molecules and relevant biological processes.

Ignore relationships that are not biological in nature (e.g., social, historical, metaphorical, or unrelated contextual information).

If you are unsure whether a relation is biological, only include it if it can plausibly describe a biological interaction or process.

The context of the text is/are: Theme: Human Immune System Description: The human immune system is a complex network of cellular and molecular interactions that works to protect the body against infectious agents and maintain homeostasis. Various biological components act together, including specialized cells that recognize and respond to external threats, as well as molecules that regulate and amplify immune responses. Understanding these interactions is essential for developing therapies to treat infections, autoimmune diseases, and other immunityrelated conditions.

Extract the relations from the given text and provide the results in JSON format. Each result must contain the following fields: - source: the source entity of the relation. It must never be null. - relation: a verb that represents the identified relation. It must never be null. - destiny: the destiny entity of the relation. It must never be null. - condition: try to infer the necessary condition for the relation to occur.

If the same concept appears with slight variations (e.g., "artificial intelligence"and "AI"), use the most common or canonical form consistently. IMPORTANT: Respond ONLY with a valid JSON object. Do not include explanations, notes, or text outside the JSON. Ensure the JSON is syntactically correct and fully closed.

Use the examples below as a guide:

Example 1

Context: "To produce antibodies, B cells must first be activated. B cells that have never been activated by encountering their cognate antigen are called naive or virgin B cells. An example would be a B cell that can recognize the smallpox virus, but which happens to reside in a human who has never been exposed to smallpox. In contrast, B cells that have encountered their cognate antigen and have been activated are called experienced B cells. There are two ways that naive B cells can be activated to defend

against invaders. One is completely dependent on the assistance of helper T cells (T cell-dependent activation) and the second is more or less independent of T cell help (T cell-independent activation)."

Expected Output (JSON format):


```
          {
"relations": [
    {
        "source": "B cells",
        "relation": "produce",
        "destiny": "antibodies",
        "condition": "after activation"
    },
    {
        "source": "naive B cells",
        "relation": "also_called",
        "destiny": "virgin B cells",
        "condition": "before encountering cognate antigen"
    },
    {
        "source": "naive B cells",
        "relation": "type_of",
        "destiny": "B cells",
        "condition": "before activation"
    },
    {
        "source": "experienced B cells",
        "relation": "type_of",
        "destiny": "B cells",
        "condition": "after antigen encounter and activation"
    },
    {
        "source": "naive B cells",
        "relation": "differentiate_into",
        "destiny": "experienced B cells",
        "condition": "after activation during immune response"
    },
    {
        "source": "B cell",
```

```
        "relation": "recognizes",
        "destiny": "smallpox virus",
        "condition": "if specific for smallpox antigen"
    },
    {
        "source": "B cell",
        "relation": "resides_in",
        "destiny": "human",
        "condition": ""
    },
    {
        "source": "human",
        "relation": "exposed_to",
        "destiny": "smallpox virus",
        "condition": ""
    },
    {
        "source": "T cell-dependent activation",
        "relation": "type_of",
        "destiny": "activation of B cells",
        "condition": ""
    },
    {
        "source": "T cell-independent activation",
        "relation": "type_of",
        "destiny": "activation of B cells",
        "condition": ""
    },
    {
        "source": "T cell-dependent activation",
        "relation": "requires",
        "destiny": "helper T cell",
        "condition": "for B cell activation"
    },
    {
        "source": "T cell-independent activation",
        "relation": "does_not_require",
        "destiny": "helper T cell",
        "condition": ""
```

```
        },
        {
            "source": "naive B cells",
            "relation": "activated_by",
            "destiny": "helper T cell",
            "condition": "in T cell-dependent activation, after antigen recognition
        }
    ] }
```

Example 2

Context: "The CD8-TL recognizes intracytoplasmic antigens presented by MHC molecules class I, which are expressed by practically all nucleated cells. Cells infected by viruses and tumor cells are normally recognized by the CD8-TL.18 After adhering to the target-cells presenting an antigen associated with the MHC and adequate co-stimulation, the CD8-TL proliferate and, at a subsequent contact, can eliminate by cytotoxicity any cell that presents this specific antigen, regardless of the presence of co-stimulatory molecules. The CD8-TL induce the apoptosis in the target cell through the action of perforins and granzymes and can also lead to apoptosis through the expression of the Fas L receptor (CD95), which interact with the Fas molecule in the target cells.18 TL A small population of peripheral TL has TCR with limited diversity, consisting of chains. These cells differ from the TL, as their TCR can recognize antigens even in the absence of presentation by the MHC molecule, being then considered true sentinels of the body.23,4. The TL also present immu -nological memory and respond more vigorously at a second antigen contact. The TL exercise their effector functions in different ways, by cytotoxicity, for instance (a primary cha -racteristic of CD8-TL). Therapeutic studies have explored this cytotoxic characteristic against tumor antigens. The TL also have a auxiliary function, releasing cytokines such as INF- (Th1) or IL-4 (Th2) and can act as efficient APC, as they have a high capacity to present antigens to the TL, mediating their activation and proliferation."

Expected Output (JSON format):

```
{
  "relations": [
      {
          "source": "CD8-TL",
          "relation": "recognizes",
          "destiny": "intracytoplasmic antigens",
          "condition": "presented by MHC class I molecules"
      },
```

```
        {
            "source": "MHC class I molecules",
            "relation": "expressed_by",
            "destiny": "nucleated cells",
            "condition": ""
        },
        {
            "source": "\u03b3\u03b4 TL",
            "relation": "recognizes",
            "destiny": "antigens",
            "condition": "even in the absence of presentation by the MHC molecule"
        },
        {
            "source": "CD8-TL",
            "relation": "eliminates",
            "destiny": "target cell",
            "condition": "presence of specific antigen and previous contact"
        },
        {
            "source": "CD8-TL",
            "relation": "induces_apoptosis_in",
            "destiny": "target cell",
            "condition": "through perforins and granzymes action OR Fas L receptor
        }
    ]
}
```


Example 3

Context: "Any invader that breaches the physical barrier of skin or mucosa is greeted by the innate immune system – our second line of defense. Immunologists call this system “innate” because it is a defense that all animals just naturally seem to have. Indeed, some of the weapons of the innate immune system have been around for more than 500 million years. Let me give you an example of how this amazing innate system works."

Expected Output (JSON format):


```
            {
    "relations": [
        {
```

```
        "source": "invader",
        "relation": "breaches",
        "destiny": "physical barrier",
        "condition": "when the barrier is compromised, such as through cuts or
    },
    {
        "source": "physical barrier",
        "relation": "includes",
        "destiny": "skin",
        "condition": ""
    },
    {
        "source": "physical barrier",
        "relation": "includes",
        "destiny": "mucosa",
        "condition": ""
    },
    {
        "source": "invader",
        "relation": "exposed_to",
        "destiny": "innate immune system",
        "condition": "after breaching the physical barrier"
    },
    {
        "source": "innate immune system",
        "relation": "part_of",
        "destiny": "second line of defense",
        "condition": ""
    },
    {
        "source": "innate immune system",
        "relation": "type_of",
        "destiny": "immune defense",
        "condition": ""
    },
    {
        "source": "innate immune system",
        "relation": "present_in",
        "destiny": "animals",
```

```
            "condition": ""
        },
        {
            "source": "innate immune system",
            "relation": "evolved_in",
            "destiny": "animals",
            "condition": ""
        }
    ]
}
```

Examples end here.

Text to analyze: "Table 1 – Basic characteristics of classes of immunoglobulins Class Structure Properties IgA Dimeric and Monomeric Found in gastrointestinal, respiratory and urogenital tract mucosa. Prevents the colonization by pathogens. Also present in saliva, tears and milk. IgD Monomeric Membrane immunoglobulin. It is part of the membrane receptor of naïve B lymphocytes (BCR). IgE Monomeric Involved in allergic and parasitic processes. Its interaction with basophils and mastocytes causes histamine release. IgG Monomeric Main immunoglobulin of acquired immunity. It has the capacity to cross the placental barrier. IgM Monomeric Pentameric It is part of the membrane receptor of naïve B lymphocytes (BCR). Form found in the serum, secreted early in acquired immune response."

Your answer: