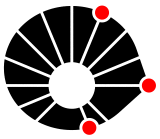

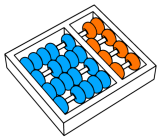

Geração de triplas RDF com base em Textos Clínicos

Thiago Luna Pinheiro *Julio Cesar dos Reis*

Relatório Técnico - IC-PFG-23-03
Projeto Final de Graduação
2023 - Dezembro

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Geração de triplas RDF com base em textos clínicos

Thiago Luna Pinheiro Julio Cesar dos Reis*

Resumo

Registros médicos, frequentemente em formatos não-estruturados como transcrições de diálogos entre médico e paciente, representam um desafio para análise, interpretação e uso efetivo no ecossistema de saúde. Há um grande desafio na transformação de textos clínicos desestruturados para dados estruturados e semanticamente ricos, como triplas RDF. Isso pode gerar melhor acesso, qualidade e organização para dados em saúde. Este estudo objetiva investigar, desenvolver e experimentar um método para geração de triplas RDF a partir de textos não-estruturados na língua Portuguesa, compostos por informações clínicas relevantes identificadas na transcrição de diálogos entre médico e paciente em consultas clínicas. Nossa proposta explorou e experimentou diversos modelos de linguagem grandes (como GPT-3 e BLOOM) e técnicas de uso dos mesmos na geração de triplas.

1 Introdução

No ambiente médico atual, os registros dos pacientes são uma fonte primária de informação. Estes registros, que contêm diagnósticos, históricos médicos, tratamentos e outras informações relevantes, são essenciais para a continuidade dos cuidados ao doente. No entanto, muitos destes registros são mantidos em formatos não-estruturados, como notas manuscritas ou transcrições de conversas. A terminologia médica, pela sua natureza, é complexa e especializada [1]. Estes factores tornam a análise e a interpretação destes registros uma tarefa altamente complexa. A falta de um padrão uniforme nos formatos de dados e a dificuldade de integração de diferentes sistemas de registros médicos aumentam o desafio, pois impedem uma visão consolidada do histórico do paciente [2], [3].

Os profissionais de saúde, sejam médicos, enfermeiros ou especialistas, ao interagirem com pacientes em consultas, coletam e atualizam constantemente informações sobre sua saúde. Cada interação é uma oportunidade para coletar dados vitais, desde sintomas relatados até resultados de exames. A capacidade de assimilar, interpretar

*Instituto de Computação, Universidade Estadual de Campinas, 13081-970 Campinas, SP.

e sintetizar essas informações é fundamental para determinar o próximo passo no tratamento do paciente [4].

Com a crescente quantidade de informação e a necessidade de uma estrutura padronizada para a armazenar os dados da saúde, os profissionais precisam de ajuda para tomar decisões informadas. Existe uma procura urgente de ferramentas e métodos de tecnologia digital que possam organizar, estruturar e visualizar estes dados de forma clara e eficaz. A digitalização avançada dos registros médicos, embora promissora, trouxe à tona desafios adicionais na extração de informações vitais de textos não-estruturados. O Processamento de Linguagem Natural (NLP) surge como uma solução potencial, prometendo revolucionar a forma como os dados são extraídos e interpretados [5]. No entanto, sua aplicação em contextos clínicos complexos não está isenta de limitações, especialmente quando se trata de identificar construções complexas, como por exemplo, discussões sobre metas de cuidado [6].

Enquanto a integração de tecnologias como o NLP promete uma revolução na gestão de dados clínicos, a jornada para sua implementação eficaz está repleta de desafios. A busca por uma solução que equilibre eficiência, precisão e acessibilidade define o contexto atual da gestão de informações clínicas. O emprego do NLP na análise de Registro de Saúde Eletrônico em formato de texto não-estruturado oferece oportunidades significativas para quantificar desfechos que, tradicionalmente, demandariam uma abstração meticulosa e onerosa dos registros médicos. Nesse contexto é notório que diversos procedimentos estatísticos frequentemente adotados em pesquisas clínicas tendem a desconsiderar potenciais erros de classificação. A aplicação desses métodos a desfechos que não foram mensurados com precisão pode resultar em estudos com poder estatístico insuficiente e estimativas distorcidas.

Uma das soluções emergentes para enfrentar os desafios da gestão de informações médicas é a utilização de Grafos de Conhecimento (do inglês - *knowledge graphs*) [7]–[9]. Os Knowledge Graphs (KGs) são estruturas de dados que permitem representar informações e suas inter-relações de forma sistemática. Eles são formados por entidades (nós) e relações (arestas) que conectam essas entidades. O Resource Description Framework (RDF) é um padrão utilizado para descrever esses dados de forma interoperável na Web[10]. As triplas RDF, que são a base deste framework, descrevem a relação entre duas entidades e consistem em três componentes: sujeito, predicado e objeto. Essa estrutura de triplas RDF é usada para modelar o relacionamento de entidades no grafo de conhecimento de forma padronizada, facilitando a integração e consulta de dados[11], assim como uma representação semanticamente rica.

O principal objetivo deste estudo é explorar e desenvolver um método que combine técnicas avançadas de NLP para estruturar dados clínicos explorando KG. Especificamente, o foco é gerar triplas RDF a partir de diálogos clínicos, como conversas entre médico e paciente. O estudo procura transformar a forma como a informação clínica é representada e armazenada, tornando-a mais organizada, acessível e fácil de interpretar e consultar. Os resultados deste estudo podem beneficiar tanto os profissionais

de saúde como os pacientes.

Desenvolvemos um método para a extração automatizada de triplas RDF. Em síntese, nossa solução determina via uma classificação do texto de entrada, se ele é relevante do ponto de vista clínico. Textos classificados como relevantes são então processados para extração de triplas. Diversos modelos e técnicas foram averiguadas para refinar a precisão e eficácia da extração. Introduzimos uma etapa adicional de sumarização, que busca condensar as informações do texto de entrada, focando nos pontos mais relevantes, antes de prosseguir para a extração de triplas.

A metodologia do estudo foi segmentada em cinco etapas principais. Na primeira etapa realiza-se o pré-processamento e adequação dos textos clínicos para análise, incluindo remoção de caracteres especiais, tokenização e segmentação. A segunda etapa sumariza o texto clínico com base no modelo *Fine-tuned GPT-NeoX 20B*. Na terceira etapa exploramos o modelo *Fine-tuned GPT-NeoX 20B* para determinar a relevância clínica dos textos. Na quarta etapa trata-se da extração de triplas com a geração de triplas RDF a partir dos segmentos relevantes dos textos, usando modelos como *GPT-3* e *BLOOM*. Na quinta etapa apresentamos a geração de triplas com aplicação da técnica de Named Entity Recognition (NER) para identificar entidades específicas e mapeá-las em triplas RDF.

A metodologia proposta mostrou-se promissora, melhorando a organização e visualização dos dados clínicos triplificados. A avaliação da classificação de diálogos médico-paciente indicou a efetividade da solução, especialmente com diálogos sumarizados.

Este documento será organizado da seguinte maneira: Na Seção 2 apresentamos uma síntese da literatura, fornecendo exemplos de trabalhos que estão relacionados ao nosso estudo; A Seção 3 descreve a metodologia de nosso estudo, apresentando o conjunto de dados e modelos utilizados, além de outros detalhes da implementação da solução e as métricas de avaliação utilizadas; A Seção 4 relata os resultados obtidos; Na Seção 5 discutimos e analisamos acerca dos resultados obtidos; Por fim, a Seção 6 reporta as conclusões do trabalho.

2 Síntese da Literatura

Nesta seção abordamos trabalhos que estão relacionados à problemática de nosso estudo. A Tabela 1 apresenta os estudos mencionados, descritos a seguir:

Rossanez e dos Reis [11] desenvolveram em seu trabalho um método para geração de KGs a partir do processamento de textos não-estruturados. Tais textos são compostos por artigos científicos do domínio da doença de Alzheimer. É realizado um pré-processamento dos dados utilizando técnicas NLP, em que estes são limpos, normalizados e transformados para facilitar sua manipulação. Então, após identificar as informações relevantes, estas são extraídas na forma de triplas RDF. E por fim,

os conceitos identificados nas triplas extraídas são vinculados a uma ontologia de domínio público.

Regino *et al.* [12] criaram o QART (Question and Answer to RDF Triples Framework), um framework que utiliza uma abordagem baseada em NLP para geração de triplas RDF a partir de Q&A de produtos de ecommerce. Visando gerar triplas precisas, confiáveis e de maneira escalável para que estas então possam contribuir no povoamento de KGs. O framework é dividido em três etapas: identificação de intenção e entidades relevantes no contexto de ecommerce; normalização e sumarização do texto original; geração de triplas RDF a partir do texto sumarizado.

Juric *et al.* [13] realizaram um estudo acerca da extração de relações entre entidades médicas. Foi então proposta uma *pipeline* de extração de informação a partir de textos não-estruturados que seria então utilizada para enriquecer bases de conhecimento. A abordadem escolhida é baseada em Anotação de Papéis Semânticos, em inglês *Semantic Role Labeling*, uma importante tarefa de Processamento de Linguagem Natural em que palavras e trechos em uma sentença recebem uma *label* que indica o seu papel semântico na sentença. Por se tratar de um método com propósito geral, modificações foram propostas ao SLR para melhorar a sua efetividade e facilitar a extração de triplas semânticas.

Tabela 1: Estudos relacionados ao processo *text-to-triple* para formação de *Knowledge Graphs*.

Ano	**Título**
2019	Generating Knowledge Graphs from Scientific Literature of Degenerative Diseases [11].
2020	A System for Medical Information Extraction and Verification from Unstructured Text[13].
2022	From Natural Language Texts to RDF Triples: A Novel Approach to Generating E-commerce Knowledge Graphs[12].

3 Metodologia

Esta seção apresenta em detalhes a metodologia aplicada para realização do processo de extração de triplas, em que dada uma entrada de texto em Linguagem Natural (na língua Portuguesa) são extraídas as triplas RDF consideradas relevantes. Em nosso caso, nossa entrada se caracteriza como diálogos médico-paciente. Para a realização de tal processo foi desenvolvido duas *pipelines* que por sua maior parte possuem estruturas idênticas, diferindo apenas no segmento inicial, em que uma delas possui um componente de sumarização.

3.1 Visão geral

A nossa metodologia está estruturada em quatro etapas principais. Na primeira etapa, inicia-se com o desenvolvimento do *dataset*, meticulosamente construído, incorporando diálogos entre profissionais médicos e pacientes, juntamente com outras informações de relevância para análises subsequentes. Na segunda etapa representa a Sumarização, nesta fase condensamos informações, focando em extrair a essência dos dados e testando diferentes formatos de entrada para otimizar o processamento subsequente. Na terceira etapa, com a classificação, os textos são categorizados com base em sua relevância clínica, garantindo que apenas informações pertinentes sejam avançadas para análise mais aprofundada, sendo este o tipo de texto apto a passar pelo processo de extração de triplas. Por último, na quarta etapa é realizado a extração de triplas, em que identificamos e conectamos entidades relevantes nos textos, formando triplas que capturam relações significativas entre os dados. Na Figura 1 temos um *overview* de nossa *pipeline*, apresentando todas as etapas de nosso processo, uma breve descrição das mesmas e seus respectivos *inputs* e *outputs*. A Figura 2 apresenta um exemplo de um dos diálogos de nosso *dataset* sendo submetido a *pipeline* completa de extração de triplas semânticas.

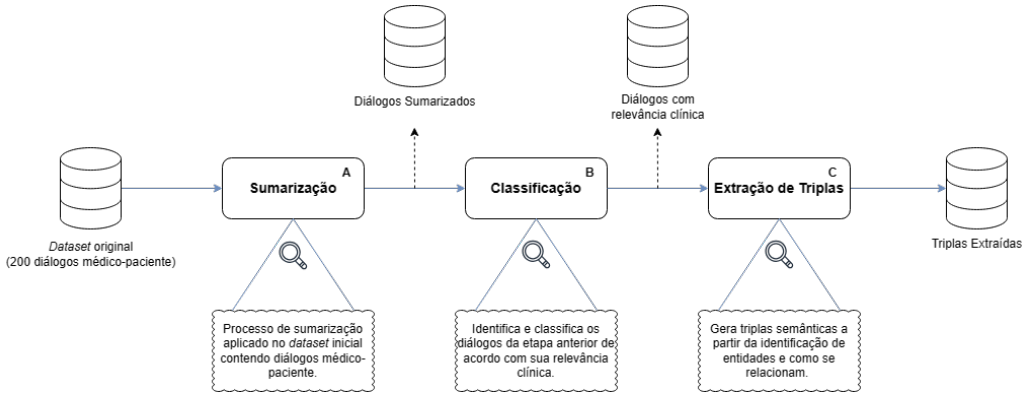

Figura 1: *Pipeline* de extração de triplas semânticas. O *pipeline* é composto por três etapas, representadas pelos retângulos ao centro da figura; na parte inferior da figura temos uma breve descrição de cada etapa; em seu entorno, *input*/*output* de cada processo.

Dentre outros aspectos relevantes que serão discutidos nesta seção podemos apontar os modelos e detalhes de implementação de cada etapa, assim como o *dataset* utilizado para realização de nossos experimentos. Já em relação a nossa segunda *pipeline*, mencionada anteriormente, temos o exato mesmo processo descrito acima sendo realizado exceto pela parte de sumarização, analisando a Figura 1 nota-se que este seria o equivalente a remoção da Etapa A.

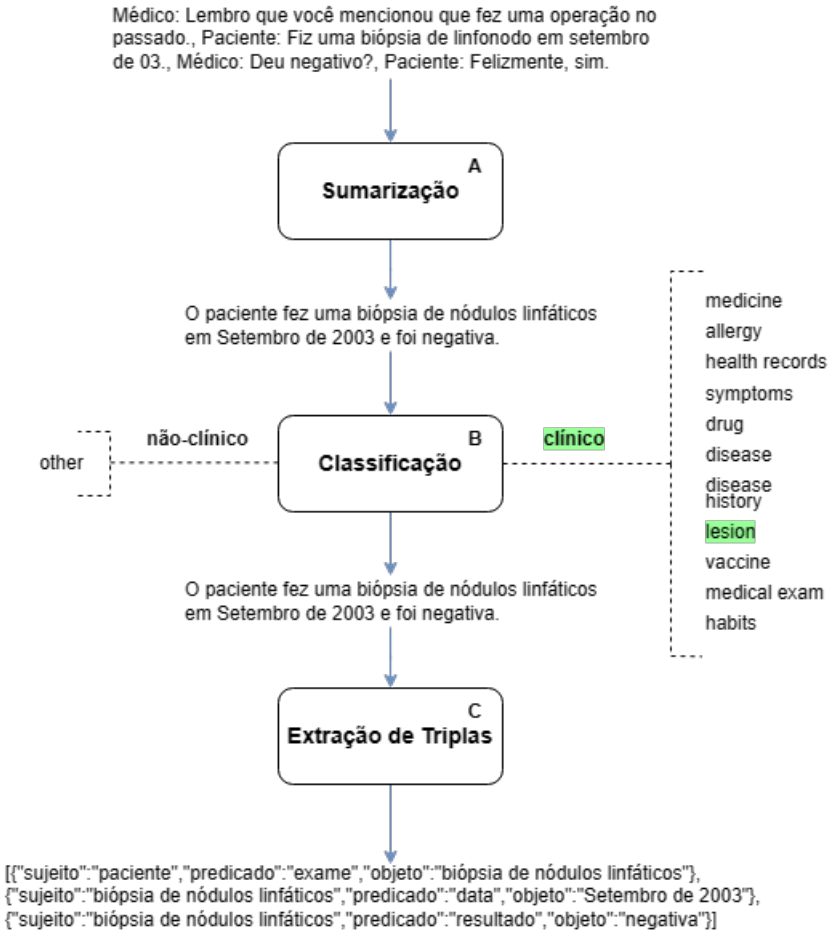

Figura 2: Exemplo de diálogo submetido à nossa proposta de extração de triplas semânticas.

3.2 Conjunto de dados

A fase inicial deste estudo envolveu a coleta de dados a partir de uma versão traduzida para o português brasileiro do *dataset* MTS-Dialog[1]. Este conjunto de dados abrange 1700 segmentos de diálogos entre médicos e pacientes e foi originado de uma pesquisa voltada para a geração de anotações clínicas a partir de interações médico-paciente.

Durante esta etapa, além da tradução, houve a necessidade de condensar o *dataset*. Assim, uma seleção aleatória foi conduzida, isolando inicialmente 200 diálogos distintos, que variavam em tamanho e temática. Este subconjunto serviu como base para a fase preliminar de análise. Posteriormente, uma subsequente seleção aleatória reduziu a amostra inicial de 200 diálogos para 20 diálogos.

O conjunto inicial de 200 diálogos foi empregado primordialmente na etapa de classificação do conteúdo como clínico ou não-clínico. Devido à natureza direta desta fase, optou-se por analisar um volume maior de diálogos, visando obter uma compreensão robusta da eficiência do processo de classificação. Em contraste, o subconjunto de 20 diálogos foi submetido integralmente à nossa *pipeline* de extração.

3.3 Sumarização

A sumarização é uma técnica que busca condensar informações de textos extensos, retendo apenas as partes mais cruciais. No contexto da "Geração de triplas RDF com base em Textos Clínicos", essa técnica é de suma importância, especialmente devido à complexidade inerente aos textos clínicos. O objetivo é simplificar a terminologia médica, facilitar a identificação e extração de relações entre entidades, otimizar o processamento de dados e, ao mesmo tempo, preservar a essência informativa do conteúdo.

O tipo de entrada utilizada em nossos modelos foi também um fator determinante na qualidade de triplas geradas/extraídas. O estudo em questão propôs duas abordagens principais para a entrada de dados nos modelos: uma que utiliza diretamente o conteúdo do *dataset* MTS, que transcreve diálogos entre médicos e pacientes, e outra que processa essa entrada para sumarizá-la, removendo coloquialismos, vícios de linguagem e outras características presentes na oralidade, tornando o texto mais direto e simplificado. Esta segunda hipótese foi levantada partindo do pressuposto de que o processo de sumarização facilitaria a extração de triplas[12].

[1] https://github.com/abachaa/MTS-Dialog

Tabela 2: Exemplificação dos tipos de entradas de nosso estudo.

Versão	Entrada
Original	Médico: Você fez alguma cirurgia no passado?, Paciente: Sim, eu fiz uma grande cirurgia de trauma há algum tempo.
Sumarizada	O médico perguntou ao paciente se ele já tinha sido operado e o paciente respondeu que já tinha sido operado de trauma grave. Graphs.

Na Tabela 2 temos uma representação do formato das entradas mencionadas anteriormente. Podemos notar a presença de dois tipos de interlocutores na versão original em questão, sendo eles: "Médico"e "Paciente". Embora estejam presentes na maior parte dos diálogos eles não são os únicos tipos de interlocutores, além deles também estão presentes em nosso *dataset*: "Médico Auxiliar"e "Convidado família". Os diferentes tipos de interlocutores não afetaram em como o processamento dos diálogos foi feito, sendo apenas um detalhe da formação de nosssas entradas.

Para a realização do processo de sumarização dos trechos de nosso *dataset* foi utilizado a plataforma NLP Cloud[2]. Nela temos acesso a um *endpoint* de *Summarization*[3], o qual foi usado para nossa tarefa. Dentre os modelos disponíveis, optamos por utilizar o Fine-tuned GPT-NeoX 20B por ser o que melhor se enquadrava com nossas necessidades, tendo bons resultados iniciais e também nativamente entender outras línguas além de inglês, em nosso caso português brasileiro.

3.4 Classificação de textos clinicamente relevantes

A classificação desempenha um papel crucial no desenvolvimento do estudo. O objetivo principal desta etapa é discernir a relevância de um texto médico, determinando se ele deve avançar para a subsequente extração de triplas. Em nosso contexto, a relevância que estamos considerando é uma relevância clínica, no caso, qualquer informação que possa contribuir para o perfil clínico do paciente que está sendo avaliado. Essa classificação é realizada por meio da identificação de entidades relevantes, que são termos ou conceitos específicos que representam informações importantes para a geração de triplas RDF. Dentre tais informações podemos citar medicamentos, sintomas, exames realizados, históricos médico e familiar, lesões, hábitos, entre outras.

Para a realização do processo de classificação foi utilizado o *endpoint Classification*[4] da plataforma NLPCloud. Assim como na etapa de sumarização, também optamos pela utilização do modelo Fine-tuned GPT-NeoX 20B.

Após determinado o nosso objetivo para a identificação do texto relevante, o que consideramos ser um texto relevante e também o que utilizariamos para realizar tal

[2]https://nlpcloud.com/
[3]https://docs.nlpcloud.com/#summarization
[4]https://docs.nlpcloud.com/#classification

classificação, iniciou-se então o estudo de qual seria a melhor maneira de separar textos clínicos e não-clínicos.

Com a utilização do modelo Fine-tuned GPT-NeoX 20B, tivemos duas opções iniciais de como realizar a classificação:

1. fornecer uma lista de *labels* (etiquetas) candidatas e então o modelo ficar a cargo de selecionar a *label* que melhor se encaixa com o texto em questão;

2. não fornecer nenhuma *label* e deixar com que o modelo classifique de forma autônoma, ou seja, uma *label* gerada pelo próprio modelo.

Decidimos seguir com a primeira alternativa para garantir respostas padronizadas, o que não seria o caso se adotássemos a segunda opção, dificultando a manipulação da resposta obtida. Seguindo a estratégia da primeira opção, foi então necessário determinar qual seria o conjunto de *labels* ideal para classificar a nossa entrada. A abordagem inicial foi de realizar uma classificação utilizando *labels* binárias, foram realizados testes fazendo a utilização de *labels* como "clínico"e "não-clínico"e outras variações neste estilo, como também "clinico"e *label* genérica. Foram obtidos resultados interessantes mas não estavamos tendo muita consistência, o que nos levou a considerar uma nova abordagem com a utilização de um conjunto maior de *labels*.

Nessa nova abordagem, a estratégia foi de incluir uma lista de diversas *labels* clínicas distintas, cada uma delas fazendo referência a um assunto clínico diferente, e então uma *label* genérica, para categorizar textos não-clínicos. Isso resultou em uma melhoria significativa na consistência dos resultados, o que nos levou a seguir com esta estratégia. Um último detalhe relevante desta parte de classificação foi a substituição da palavra "médico"antes de submeter o texto pelo processo de classificação, tal decisão foi tomada visando evitar que textos que não tivessem uma relevância clínica fossem classificados como tal apenas por possuírem a palavra "médico".

Para que fosse possível avaliar a efetividade do modelo e da estratégia adotada na tarefa de classificação foi criado um *dataset* padrão-ouro, contendo a classificação esperada de 200 diálogos, mencionados na Seção 3.2. Este processo foi aplicado tanto na versão original quanto na sumarizada dos diálogos. Para avaliação foram utilizadas métricas como acurácia, precisão, *recall* e F_1-*score*, que serão discutidas posteriormente.

3.5 Extração de triplas

A extração de triplas é uma etapa crucial após o pré-processamento do texto. Esta fase envolve a transformação do texto de entrada em um conjunto de triplas, passando por duas sub-etapas específicas. Para entender o conceito, considere uma sentença simples $s = \{np, vp\}$, composta de um sintagma nominal e um sintagma verbal. O sintagma verbal, por sua vez, pode ser composto por um verbo v, (o verbo principal da

frase) e outro sintagma nominal, ou seja, vp = $\{v, np'\}$. Tal verbo no sintagma verbal denota uma relação entre os dois sintagmas nominais, considerados os argumentos verbais. Essa relação pode ser representada por um tripla t = (s, p, o), onde o predicado se refere ao verbo, o sujeito é o primeiro sintagma nominal e o objeto, o segundo sintagma nominal, ou seja, t = (np, v, np'). Esta tripla marca a principal relação da frase.

A extração de triplas é uma técnica que permite representar as relações semânticas presentes em um texto de forma estruturada. No contexto desse estudo, essa tarefa é desempenhada usando uma combinação de modelos e estratégias. Para realizar essa tarefa, propomos a utilização e experimentação de diferentes abordagens. 1) Uso do modelo REBEL[5], uma variação multilíngue do modelo REBEL. O mREBEL é um modelo de extração de relação entre entidades que foi treinado em um *dataset* multilíngue. Ele reformula a tarefa de extração de relações como uma tarefa *seq2seq*, que é um tipo de modelo usado para predição de sequências. Dada a sua capacidade multilíngue e sua abordagem *seq2seq*, o mREBEL foi considerado uma escolha para a pipeline desta proposta. 2) sA utilização de modelos de geração de texto, como o Fine-tuned GPT-NeoX e o BLOOM[6] para a tarefa de extração de triplas. Esses modelos são usados para complementar e aprimorar o processo de extração de triplas, garantindo que todas as relações relevantes sejam identificadas e representadas de forma adequada.

3.5.1 mREBEL

O mREBEL, conforme destacado anteriormente, é um modelo *open source* especializado na extração de relações entre entidades. Sua formação foi baseada em um *dataset* multilíngue, o que permitiu que ele adaptasse a tarefa de extração de relações para um formato *seq2seq*. Esse tipo de modelo é comumente empregado em tarefas que envolvem predições sequenciais, como sumarização de texto e tradução automática.

A versatilidade multilíngue do mREBEL, combinada com sua habilidade em extrair e interligar entidades, o tornou particularmente adequado para as demandas do nosso projeto. Por isso, ele foi escolhido como um componente essencial da nossa *pipeline*.

O código a seguir representa o código utilizado em nossa *pipeline* para testes com o modelo em questão:

```
from transformers import AutoModelForSeq2SeqLM, AutoTokenizer

def extract_triplets_typed(text):
    triplets = []
    relation = ''
    text = text.strip()
```

[5]https://huggingface.co/Babelscape/mrebel-large
[6]https://huggingface.co/bigscience/bloom

```
    current = 'x'
    subject, relation, object_, object_type, subject_type = '','','','','',''

    for token in text.replace("<s>", "").replace("<pad>", "").replace("</s>", "").replace("tp_XX", "").replace("__en__", "").split():
        if token == "<triplet>" or token == "<relation>":
            current = 't'
            if relation != '':
                triplets.append({'head': subject.strip(), 'head_type': subject_type, 'type': relation.strip(),'tail': object_.strip(), 'tail_type': object_type})
                relation = ''
            subject = ''
        elif token.startswith("<") and token.endswith(">"):
            if current == 't' or current == 'o':
                current = 's'
                if relation != '':
                    triplets.append({'head': subject.strip(), 'head_type': subject_type, 'type': relation.strip(),'tail': object_.strip(), 'tail_type': object_type})
                object_ = ''
                subject_type = token[1:-1]
            else:
                current = 'o'
                object_type = token[1:-1]
                relation = ''
        else:
            if current == 't':
                subject += ' ' + token
            elif current == 's':
                object_ += ' ' + token
            elif current == 'o':
                relation += ' ' + token
    if subject != '' and relation != '' and object_ != '' and object_type != '' and subject_type != '':
        triplets.append({'head': subject.strip(), 'head_type': subject_type, 'type': relation.strip(),'tail': object_.strip(), 'tail_type': object_type})
    return triplets

# Load model and tokenizer
tokenizer = AutoTokenizer.from_pretrained("Babelscape/mrebel-large", src_lang="pt_XX", tgt_lang="tp_XX")

model = AutoModelForSeq2SeqLM.from_pretrained("Babelscape/mrebel-large")
gen_kwargs = {
```

```
    "max_length": 256,
    "length_penalty": 0,
    "num_beams": 8,
    "num_return_sequences": 3,
    "forced_bos_token_id": None,
}
```

O código apresentado utiliza a biblioteca Hugging Face Transformers para carregar e operar um modelo *seq2seq*, especificamente o "Babelscape/mrebel-large". A principal função, *extract_triplets_typed*, analisa um texto fornecido para identificar triplas, que consistem em duas entidades e uma relação entre elas. Cada entidade também possui um tipo associado. Para extrair essas triplas, o código passa por cada token do texto, monitorando o estado atual (se está lendo um sujeito, objeto ou relação) e usa várias tags para ajudar na identificação. Uma vez identificadas todas as partes da tripla, ela é adicionada a uma lista de triplas. Ao final do processo de extração, um tokenizador e um modelo são carregados, e são definidos parâmetros específicos para a geração subsequente de sequências pelo modelo.

O código base, assim como outras informações sobre o modelo, pode ser encontrado na plataforma Hugging Face. O modelo mREBEL, do qual uma versão é usada nesse estudo, possui diversas versões, sendo a *large* a mais extensa, capaz de detectar mais de 400 tipos de relações entre entidades.

3.5.2 Técnica explorando Modelos de Geração de Texto

Os modelos de geração de texto são sistemas de aprendizado de máquina treinados para gerar sequências de texto coerentes e contextualmente relevantes. Com avanços recentes na área de processamento de linguagem natural, esses modelos tornaram-se instrumentos vitais em tarefas que exigem compreensão e geração de linguagem.

Seleção de Modelos Antes de começar a experimentação, foi feita uma avaliação cuidadosa dos modelos disponíveis no mercado. Após uma análise detalhada, optou-se pelo Fine-tuned GPT-NeoX, por meio do *endpoint* de *Text Generation*[7] na plataforma NLP Cloud, e pelo BLOOM, disponível na plataforma Hugging Face. A escolha foi baseada na eficácia, capacidade de customização e acessibilidade dos modelos.

Definição de *Prompts* Antes de executar a tarefa de geração, foi crucial definir como orientar os modelos para a tarefa em questão. Tal ato de fornecer um contexto e instruir o modelo em como realizar certa tarefa é realizado por meio de um *prompt*. Para resolução de nosso problema, foram adotadas três estratégias diferentes para o desenvolvimento destes *prompts*:

1. ***few-shot prompting*:** o modelo é alimentado com alguns exemplos de entrada e as respectivas saídas esperadas. Em nosso caso foi fornecido um conjunto de

[7] https://docs.nlpcloud.com/#generation

15 frases clínicas e então como saída espearada as triplas extraídas para cada uma dessas frases;

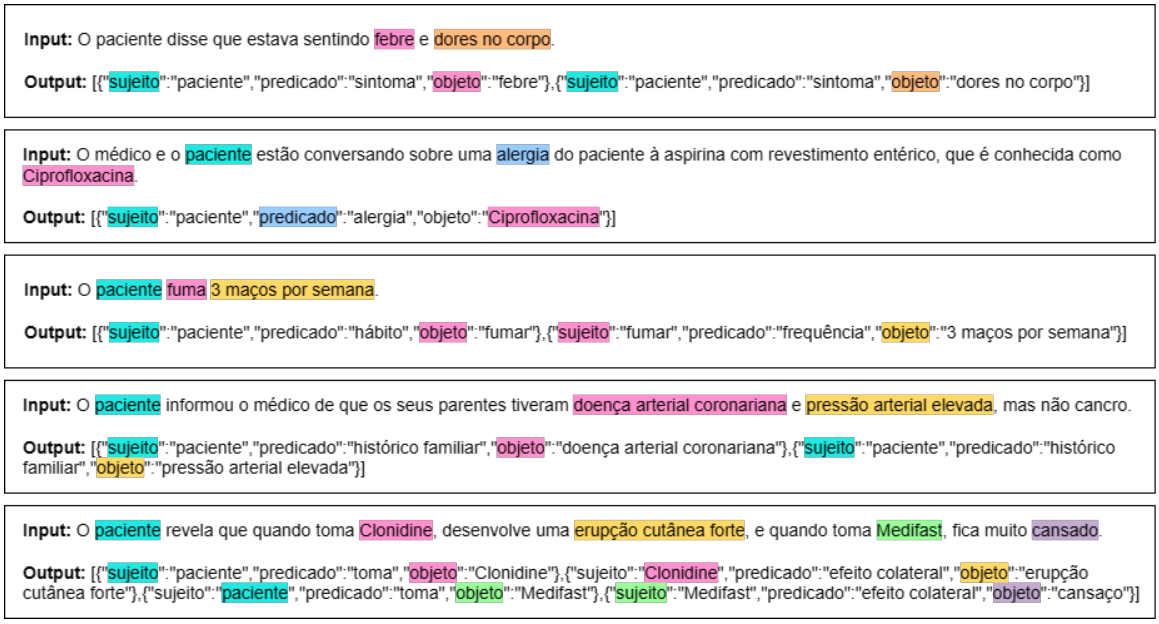

Figura 3: Exemplos de *few-shot prompting* utilizados. Pode-se notar que há casos em que o objeto também exerce papel de sujeito, assim como alguns casos em que o predicado é inferido dependendo do contexto da sentença.

2. **instrução simples:** por meio de linguagem humana natural é dado uma instrução clara e concisa ao modelo de como realizar a tarefa em questão. Em nosso caso foi socilitado para que fossem extraídas as triplas semânticas dada uma entrada e que o resultado fosse fornecido no formato JSON;

3. **combinação de ambas:** nesta abordagem foi realizada uma combinação de ambas estratégias utilizadas anteriormente, resultando então em um *prompt* com uma instrução em linguagem natural em seu início seguido de alguns exemplos no estilo de *few-shot learning*.

Parâmetros. Depois de definir os prompts, a próxima etapa foi ajustar os parâmetros de geração para otimizar a qualidade da saída. Dois parâmetros fundamentais que podemos mencionar são a *temperatura* e o $top - p$.

A *temperatura* foi ajustada para controlar a aleatoriedade da saída. Seu valor for definido nos modelo Fine-tuned GPT-NeoX e BLOOM como 0 e 0.1, respectivamente, com a finalidade de obter resultados mais determinísticos, em que dada uma entrada a saída gerada se manteria a mesma, independente do número de execuções. Já em

relação ao parâmetro $top - p$, os experimentos foram feitos variando seu valor entre 0.5 e 1 para o modelo da família GPT e entre 0.1, 0.5 e 0.9 para o modelo BLOOM. Nele é determinado que apenas o conjunto de tokens mais prováveis e que tem a soma de suas probabilidades igual ou maior a p será selecionado.

Execução e Avaliação Com os modelos selecionados, dados preparados, prompts definidos e parâmetros ajustados, os modelos foram então executados. As saídas geradas foram comparadas com os resultados esperados para avaliar a eficácia dos modelos e da metodologia adotada.

Iteração Com base nos resultados obtidos, fez-se uma revisão da metodologia. Ajustes foram feitos conforme necessário, seja na seleção de modelos, definição de prompts ou parametrização, e o processo foi repetido até que os resultados desejados fossem alcançados. Após várias iterações e avaliações, chegou-se à conclusão de que, ao utilizar a metodologia correta, os modelos de geração de texto podem ser extremamente eficazes em tarefas como a extração de triplas semânticas de frases clínicas. A chave para o sucesso reside na combinação certa de seleção de modelo, definição de prompt e ajuste de parâmetros.

3.6 Métricas de Avaliação

Buscando avaliar empregamos métricas de avaliação. Para essa avaliação foram selecionadas quatro métricas principais: acurácia, *recall*, precisão e *F1-score*. A acurácia ofereceu uma visão geral de como o modelo se desempenhou. O *recall* indiciou a quantidade de casos positivos ("clínicos") classificados corretamente em relação ao total previsto. Já com a precisão é analisado a quantidade de casos corretos dentre os classificados como positivos. Finalmente, temos *F1-score* como sendo uma média harmônica entre precisão e *Recall*, proporcionando um balanço entre ambas as métricas.

$$Ac = \frac{VP + VN}{P + N} = \frac{VP + VN}{VP + VN + FP + FN} \tag{1}$$

$$Recall = \frac{VP}{P} = \frac{VP}{VP + FN} \tag{2}$$

$$Prec = \frac{VP}{VP + FP} \tag{3}$$

$$F_1 = \frac{2 * Prec * Recall}{Prec + Recall} \tag{4}$$

As equações de 1 a 4 representam como foi feito o cálculo das métricas descritas anteriormente. Na Equação 1 temos a descrição matemática da fórmula da acurácia, logo em seguida temos *recall* (Equação 2), precisão (Equação 3) e por último *F1-score* (Equação 4). Em relação as siglas utilizas e composição das equações, temos que:

- **Verdadeiro Positivo (VP):** classe Positiva classificada corretamente;
- **Verdadeiro Negativo (VN):** classe Negativa classificada corretamente;
- **Falso Positivo (FP):** classificação incorreta em que o resultado previsto é dado como Positivo mas o resultado real e esperado é Negativo;
- **Falso Negativo (FN):** classificação incorreta em que o resultado previsto é dado como Negativo mas o resultado esperado é Positivo;
- **Positivo (P):** total de todos resultados Positivos esperados;
- **Negativo (N):** total de todos resultados Negativos esperados.

4 Resultados Experimentais

A seção 4.1 apresenta os resultados sobre a classificação de textos clínicos, enquanto a seção 4.2 apresenta os resultados sobre as técnicas avaliadas para a extração de triplas RDF.

4.1 Classificação de Textos Clínicos

Inicialmente, foram analisados os resultados obtidos na etapa de classificação dos 200 diálogos médico-paciente mencionados anteriormente, considerando tanto sua versão original quanto em sua versão sumarizada. O desempenho do modelo de classificação foi avaliado comparando suas previsões com as classificações contidas em nosso *dataset* Padrão-ouro (*Gold Standard*), estabelecido previamente.

A Tabela 3 apresenta a classificação dos 200 diálogos em sua forma original, sem sumarização, e comparado com o resultado esperado (Padrão-ouro). A tabela está dividida em textos classificados como "clínico"e "não-clínico"e seus respectivos totais. A análise revela que, dos 200 diálogos, o modelo foi capaz de classificar corretamente 160 dos diálogos "clínicos"e 17 dos diálogos "não-clínicos". No entanto, houve uma discrepância em 23 diálogos, apresentando 9 casos de falso positivo e 14 casos de falso negativo. A classe Positiva, em nosso caso, é designada para textos categorizados como "clínicos", enquanto a classe Negativa está associada aos textos "não-clínicos".

Na Tabela 4 temos esse mesmo formato, porém focamos na análise dos diálogos após sua sumarização, com o intuito de avaliar a capacidade do modelo de classificação em discernir entre as categorias estabelecidas, mesmo quando submetido a informações mais condensadas. Dos 200 diálogos sumarizados analisados, 174 foram categorizados como pertencentes à classe "Positivo"e 26 diálogos à classe "Negativo".

Tabela 3: Comparação entre a classificação de nosso padrão-ouro e a prevista por nosso modelo para os 200 diálogos médico-paciente. A classificação em questão foi feita tendo os diálogos em sua forma original como entrada.

		Predição (Versão Original)		
		clínico	**não-clínico**	**total**
Padrão-ouro	**clínico**	160	14	174
	não-clínico	9	17	26
	total	169	31	200

O resultado é composto por 158 casos de Verdadeiro Positivo (VP), 18 casos de Verdadeiro Negativo (VN), 16 casos de Falso Negativo (FN) e 8 casos de Falso Positivo (FP).

Tabela 4: Comparação entre a classificação de nosso padrão-ouro e a prevista por nosso modelo para os 200 diálogos médico-paciente. A classificação em questão foi feita tendo os diálogos em sua forma sumarizada como entrada.

		Predição (Versão Sumarizada)		
		clínico	**não-clínico**	**total**
Padrão-ouro)	**clínico**	158	16	174
	não-clínico	8	18	26
	total	166	34	200

Em relação as métricas de avaliação, na Tabela 5 podemos visualizar os valores obtidos na classificação dos diálogos em sua forma original. Pela Equação 1, nota-se que 177 casos (88.5%) foram classificados corretamente do total de 200 diálogos avaliados, representando a acurácia do modelo utilizado. Além disso, temos *Recall* atingindo a marca de 92.5%, calculado utilizando a Equação 2; Precisão em 94,7%, calculado pela Equação 3 e por fim F_1 *score* alcançando 93.6% e obtido por meio da Equação 4.

Já na Tabela 6, temos a apresentação das mesmas métricas mas em relação a classificação realizada nos diálogos sumarizados. Para Acurácia, *Recall* e Precisão obtivemos 88,0%, 91,3% e 95,2%, respectivamente. Já em relação ao F_1-*score* foi obtido um valor de 93,2%. Indicando uma boa qualidade na estratégia utilizada para classificação da relevância clínica dos textos analisados, em amabas estratégias utilizadas.

Tabela 5: Resultado do cálculo das métricas de avaliação: Acurácia, *Recall*, Precisão e F_1-*score* para os 200 diálogos médico-paciente. A classificação em questão foi feita tendo os diálogos em sua forma original como entrada.

Ac	**Recall**	**Prec**	$\mathbf{F_1}$
88.5%	92.5%	94.7%	93.6%

Tabela 6: Resultado do cálculo das métricas de avaliação: Acurácia, *Recall*, Precisão e F_1-*score* para os 200 diálogos médico-paciente. A classificação em questão foi feita tendo os diálogos em sua forma sumarizada como entrada.

Ac	**Recall**	**Prec**	$\mathbf{F_1}$
88.0%	91.3%	95.2%	93.2%

4.2 Resultados de extração Triplas RDF

Para essa subseção, será apresentado os resultados obtidos com a execução da *pipeline* em sua totalidade. Nesse processo foram submetidos apenas o subconjunto de 20 diálogos médico-paciente selecionados aleatoriamente dos 200 diálogos da etapa anterior. Inicialmente, serão exibidos os resultados referentes ao processo de classificação do subconjunto em questão. Em seguida será apresentado uma analisa qualitativa das triplas geradas na etapa de extração de triplas, tendo um foco nos melhores resultados de cada modelo, entre outros aspectos.

Na Tabela 7, é realizada uma comparação da classificação prevista por nosso modelo com a classificação esperada, presente em nosso *dataset* Padrão-ouro. A tabela em questão apresenta os resultados obtidos em ambas estratégias adotadas, tanto na que utiliza o diálogo original como entrada quanto na que faz uso de sua versão sumarizada. Analisando a tabela, temos 13 e 14 casos de Verdadeiro Positivo para as versões original e sumarizada, respectivamente. Assim como 3 e 4 casos de Verdadeiro Negativo . Já em relação as classificações previstas, temos que a versão original apresentou 1 caso de Falso Positivo e 3 casos de Falso Negativo. Por outro lado, a versão sumarizada teve 0 e 2 casos destes mesmos tipos de erros, respectivamente.

Tabela 7: Comparação entre a classificação de nosso padrão-ouro e a prevista por nosso modelo para os 20 diálogos médico-paciente. A classificação foi feita tendo como entrada tanto os diálogos em sua forma original quanto sumarizada.

		Predição (Versão Original)			**Predição (Versão Sumarizada)**		
		clínico	**não-clínico**	**total**	**clínico**	**não-clínico**	**total**
Padrão-ouro	**clínico**	13	3	16	14	2	16
	não-clínico	1	3	4	0	4	4
	total	14	6	20	14	6	20

Utilizando os resultados apresentados na Tabela 7 foi calculado os valores das métricas de acurácia, *recall*, precisão e F_1-*score* de ambas abordagens. A Tabela 8 exibe os resultados referentes a estratégia que utiliza os diálogos em sua forma original como entrada. A acurácia nesse cenário foi de 80.0%. Já o *recall* e a precisão alcançaram valores de 81.3% e 92.9%, respectivamente. Por fim, o F_1-*score* atingiu 86.7%. Quanto a Tabela 9, versão sumarizada, os valores obtidos foram 90.0% e 87.5%, para acurácia e *recall*. 100% e 93.3% para precisão e F_1-*score*.

Tabela 8: Resultado do cálculo das métricas de avaliação para os 20 diálogos médico-paciente analisados: Acurácia, *Recall*, Precisão e F_1-*score*. A classificação em questão foi feita tendo os diálogos em sua forma original como entrada.

Ac	**Recall**	**Prec**	**F_1**
80.0%	81.3%	92.9%	86.7%

Tabela 9: Resultado do cálculo das métricas de avaliação para os 20 diálogos médico-paciente analisados: Acurácia, *Recall*, Precisão e F_1-*score*. A classificação em questão foi feita tendo os diálogos em sua forma sumarizada como entrada.

Ac	**Recall**	**Prec**	**F_1**
90.0%	87.5%	100.0%	93.3%

Em relação a análise qualitativa das triplas:

mREBEL. Dos três modelos avaliados em nosso estudo, o mREBEL foi o que pior desempenhou na tarefa de extração de triplas, em ambos tipos de entrada - original e sumarizada. As triplas extraídas por esse modelo não alcançaram um resultado decente, podendo ser separadas em dois casos: triplas incoerentes e confusas; e triplas que não expressam as principais informações relatadas. Dessa forma, nota-se o mREBEL falhou em gerar triplas contendo informações relevantes e que pudessem contribuir na estruturação do quadro clínico do paciente.

BLOOM. Para o modelo BLOOM, um dos modelos de geração de texto utilizados, foram realizados testes utilizando a estratégia de *few-shot prompting* em conjunto com a variação do parâmetro $top - p$. Em relação aos seus resultados, apresentou um desempenho bem melhor comparado ao modelo mREBEL, visto que foi capaz de produzir triplas de qualidade em boa parte dos testes realizados. Embora tenha apresentado bons resultados, foi possível notar uma falta de objetividade e padronização das triplas extraídas, quando utilizado como entrada a versão original dos diálogos, também estando presente em casos de entrada sumarizada. Um aspecto relevante observado foi que tais triplas por muitas vezes falharam em sintetizar as informações processadas, acarretando na formação de triplas que utilizam longos trechos da conversa para preencher um único campo, ao invés de utilizar palavras-chaves, entidades ou conceitos. O campo "objeto"foi o mais afetado por tal erro. Também tiveram casos em que o campo "objeto"não foi gerado.

Fine-tuned GPT-NeoX 20B. O modelo que melhor desempenhou em nossos testes foi o modelo da família GPT. Além da variação do parâmetro $top-p$, foram testadas três abordagens diferentes em relação ao tipo de *prompt* utilizado. Comparado com os modelos testados previamente, o modelo Fine-tuned GPT-NeoX 20B teve um desempenho superior independente da estratégia de *prompt* que estivesse utilizando.

O *prompt* formado por instrução simples em linguagem natural obteve resultados significativos porém foi o que mais apresentou problemas. Dentre eles, pode-se ressaltar a inconsistência na formatação da saída, em que, apesar de todos os resultados de nossos testes terem gerado uma saída em JSON, o modelo não foi capaz de fornecer um padrão no formato da saída, variando a maneira como esta era exibida. Além disso, quando comparado com as duas outras estratégias de *prompt*, o modelo apresentou uma dificuldade maior na tarefa de identificar informação relevante para formação da tripla.

As estratégias de *few-shot prompting* e combinação de *few-shot prompting* com instrução simples apresentaram objetivamente resultados semelhantes. Ambas foram capazes de fornecer a saída em um formato padronizado, sendo fornecido um JSON válido. As triplas extraídas apresentaram alta qualidade, contendo informações relevantes e substanciais para a composição do quadro clínico do paciente. Em relação ao tipo de entrada, o modelo teve desempenho semelhante em ambos cenários. Para o valor de $top-p$, notou-se que o modelo desempenhou melhor quando tal parâmetro teve seu valor igual a 0.5.

	Input	mREBEL	BLOOM	Fine-tuned GP-NeoX
Versão Original	Médico: Você já fez algum procedimento cirúrgico?, Paciente: Sim, na verdade eu fiz uma histerectomia em março de noventa e nove., Médico: Entendi.	[{'head': 'nove', 'head_type': 'concept', 'type': 'follows', 'tail': 'noventa', 'tail_type': 'concept'}]	[{"sujeito":"paciente","predicado": "cirurgia","objeto":"histerectomia"}, {"sujeito":"histerectomia", "predicado":"ano","objeto":"março de noventa e nove"}]	[{"sujeito":"paciente","predicado": "cirurgia","objeto":"histerectomia"}]
Versão Sumarizada	O médico perguntou à paciente se tinha feito alguma cirurgia e a paciente respondeu que tinha feito uma histerectomia em Março de 1999, ao que o médico respondeu que compreendia.	[{'head': 'histerectomia', 'head_type': 'concept', 'type': 'subclass of', 'tail': 'cirurgia', 'tail_type': 'concept'}]	[{"sujeito":"paciente","predicado": "cirurgia","objeto":"histerectomia"}, {"sujeito":"histerectomia", "predicado":"data","objeto": "Março de 1999"}]	[{"sujeito":"paciente","predicado": "cirurgia","objeto":"histerectomia"}, {"sujeito":"histerectomia", "predicado":"data","objeto": "Março de 1999"}]

Figura 4: Comparação entre os modelos utilizados para a tarefa extração de triplas.

Na Figura 4, pode-se observar as triplas extraídas por cada um dos modelos utilizados para tarefa de extração em relação a um exemplo de entrada em sua versão original e também sumarizada. O modelo BLOOM utilizou parâmetro *temperatura* igual a 0.1 e $top - p$ igual a 0.9. Já o modelo Fine-tuned GPT-NeoX utilizou os parâmetros de *temperatura* e $top - p$ a 0 e 0.5, respectivamente. Ambos utilizaram a estratégia de *few-shot prompting* para o exemplo em questão.

5 Discussão

Na análise qualitativa, o mREBEL foi menos eficaz, enquanto o BLOOM, apesar de superior, carecia de objetividade. O Fine-tuned GPT-NeoX 20B emergiu como o modelo mais efetivo, especialmente com o parâmetro $top-p$ ajustado para 0.5. Conclui-se que a abordagem proposta potencializa a acessibilidade das informações para profissionais de saúde e oferece *insights* para as comunidades de NLP e Informática em Saúde. Contudo, é crucial reconhecer que os resultados são iniciais, necessitando de mais pesquisas e avaliações para sua validação e refinamento.

A partir dos dados apresentados, é possível inferir que o modelo de classificação teve um desempenho robusto ao classificar diálogos médico-paciente tanto em sua forma original quanto sumarizada. No geral, a acurácia do modelo superou os 88% em ambos os casos, indicando que a maioria das previsões foi correta.Esta precisão indica que o modelo é altamente confiável na previsão correta da natureza dos diálogos, sejam eles descrições detalhadas ou sumarizadas de interações clínicas.

O Recall, uma métrica essencial na detecção de informações relevantes, também mostrou desempenho promissor. Com um recall de 92,5% para os diálogos originais e 91,3% para os sumarizados, o modelo tem uma alta taxa de sucesso na identificação correta dos textos clínicos reais. Essa alta taxa é fundamental no contexto médico, onde perder informações cruciais pode ter consequências significativas para diagnósticos e tratamentos.O aspecto mais intrigante, no entanto, reside na comparação entre as classificações de textos originais e sumarizados. Surpreendentemente, os resultados foram notavelmente semelhantes. Isso pode ser interpretado de duas formas:

ou a sumarização foi feita mantendo as informações mais críticas intactas, ou o modelo tem uma capacidade excepcional de deduzir o conteúdo mesmo de textos sumarizados, reiterando sua robustez.

No entanto, toda tecnologia tem espaço para melhorias. Para refinar ainda mais o processo de classificação, foi proposto expandir a lista pré-definida de *labels*, que é fornecida ao modelo responsável por escolher a que melhor se enquadra com o texto avaliado. Esta expansão tem como objetivo abranger um espectro mais amplo de temas clínicos e, assim, minimizar ainda mais a chance de erros na classificação. Consideramos esta como sendo uma possível forma de melhorar a etapa em questão considerado o avanço obtido anteriormente com o aumento no número de *labels*. Contudo, a expansão dos labels deve ser feita com cautela para manter a precisão e não sobrecarregar o modelo.

Avançando para a extração de triplas, modelos generativos, como o BLOOM e o Fine-tuned GPT-NeoX, demonstraram ser os mais eficientes, especialmente quando implementados com a estratégia de *few-shot prompting*. No entanto, um desafio persistente é a dificuldade dos modelos em diferenciar entre confirmações, negações e simples menções. Essa limitação é significativa, pois pode resultar na extração de triplas que não refletem corretamente a realidade do diálogo, levando a possíveis interpretações errôneas em contextos clínicos.

Versão	Entrada	Triplas Extraídas
O	Médico: Você tem experimentado alguma dificuldade mental ou confusão?, Paciente: Não., Médico: Alguma alucinação?,Paciente: **Não**.	[{"sujeito":"paciente","predicado":"sintoma","objeto":"dificuldade mental"}, {"sujeito":"paciente","predicado":"sintoma","objeto":"confusão"}, {"sujeito":"paciente","predicado":"sintoma","objeto":"alucinações"}]
S	O médico perguntou ao paciente se eles tinham alguma dificuldade mental, confusão ou alucinações, e o paciente disse **não** a tudo isso.	[{"sujeito":"paciente","predicado":"sintoma","objeto":"dificuldade mental"}, {"sujeito":"paciente","predicado":"sintoma","objeto":"confusão"}, {"sujeito":"paciente","predicado":"sintoma","objeto":"alucinação"}]

Figura 5: Exemplificação de uma possível incosistência na extração de triplas. Versão "O"faz referência a versão original dos diálogos, ao passo que versão "S"refere a sua versão sumarizada.

Na Figura 5 é apresentado como essa inconsistência pode ocorrer. No exemplo em questão a tripla foi extraída utilizando o modelo Fine-tuned GPT-NeoX 20B, com uso de *few-shot prompting* e parâmetro $top-p$ em 0.5. Como mencionado anteriormente, o problema esteve presente em todos os testes realizados com a entrada em questão. A variação dos modelos utilizados, bem como suas diferentes estratégias de *prompts* e alteração de parâmetros, no caso dos modelos generativos, não demonstrou ser suficiente para evitar a inconsistência em questão. O resultado ideal e esperado para uma inconsistência dessa natureza seria gerar triplas que sejam condizentes com o que foi enunciado ou não extrair nenhuma tripla, visto que as informações apresentadas não fariam parte do quadro clínico do paciente. No caso do exemplo

exibido, triplas que expressem que o paciente não apresentou dificuldade mental, confusão e alucinação ou nenhuma.

Para aprimorar e refinar a *pipeline* atual de forma a evitar a ocorrência de um problema dessa natureza, poderíamos incluir uma etapa de validação após a extração de triplas. Nesta nova fase seria verificado se as triplas extraídas condizem com os fatos enunciados no texto de entrada. Além disso, outra possível estratégia para melhorar os resultados obtidos seria a utilização de outros modelos generativos, como por exemplo, o modelo GPT-4.

6 Conclusão

Neste estudo, exploramos a interseção de técnicas avançadas de NLP e KG com o objetivo de aprimorar a representação, interpretação e organização de informações clínicas. Através da criação de um método de extração de informação, geramos triplas RDF a partir da transcrição de diálogos clínicos, especificamente conversas entre médicos e pacientes. Os resultados obtidos retratam uma evolução na forma de expressar, representar e armazenar informações clínicas. Nossos resultados demonstraram que a introdução de uma etapa de sumarização do texto, responsável por sintetizar e destacar os pontos de maior relevância se mostrou de grande valor ao processo de extração de triplas. Ao avaliar diferente modelos e abordagens, nosso estudo revelou um desempenho notável nas tarefas de extração, categorização e triplificação de dados clínicos. Concluímos que este trabalho não apenas válida a efetividade da abordagem proposta, como também pavimenta um caminho promissor para aprimorar a gestão, manipulação e interpretação de registros clínicos, com potencial impacto significativo para a área da saúde.

Agradecimentos

Este trabalho foi conduzido em colaboração com o projeto "Aplicação Interativa para Captura e Refinamento de Informação Clínica" – convênio 93943 – entre Unicamp e a Precision Data Engenharia e Ciência de Dados LTDA. Agradecemos toda a colaboração dos envolvidos na empresa no contexto desse projeto.

Referências

[1] A. Kormilitzin, N. Vaci, Q. Liu e A. Nevado-Holgado, "Med7: A transferable clinical natural language processing model for electronic health records," *Artificial Intelligence in Medicine*, v. 118, p. 102 086, 2021.

[2] S. Wu, K. Roberts, S. Datta et al., "Deep learning in clinical natural language processing: a methodical review," *Journal of the American Medical Informatics Association*, v. 27, n. 3, pp. 457–470, 2020.

[3] M. Honnibal e M. Johnson, "An improved non-monotonic transition system for dependency parsing," em *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 1373–1378.

[4] S. Velupillai, H. Suominen, M. Liakata et al., "Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances," *Journal of biomedical informatics*, v. 88, pp. 11–19, 2018.

[5] J. Sarzynska-Wawer, A. Wawer, A. Pawlak et al., "Detecting formal thought disorder by deep contextualized word representations," *Psychiatry Research*, v. 304, p. 114 135, 2021.

[6] G. B. Melton e G. Hripcsak, "Automated detection of adverse events using natural language processing of discharge summaries," *Journal of the American Medical Informatics Association*, v. 12, n. 4, pp. 448–457, 2005.

[7] M. R. Kamdar e M. Dumontier, "An Ebola virus-centered knowledge base," *Database*, v. 2015, bav049, 2015.

[8] S. Kanza e J. G. Frey, "A new wave of innovation in Semantic web tools for drug discovery," *Expert Opinion on Drug Discovery*, v. 14, n. 5, pp. 433–444, 2019.

[9] T. Ruan, Y. Huang, X. Liu, Y. Xia e J. Gao, "QAnalysis: a question-answer driven analytic tool on knowledge graphs for leveraging electronic medical records for clinical research," *BMC medical informatics and decision making*, v. 19, pp. 1–13, 2019.

[10] K. S. Candan, H. Liu e R. Suvarna, "Resource description framework: metadata and its applications," *Acm Sigkdd Explorations Newsletter*, v. 3, n. 1, pp. 6–19, 2001.

[11] A. Rossanez e J. C. dos Reis, "Generating Knowledge Graphs from Scientific Literature of Degenerative Diseases.," em *SEPDA@ ISWC*, 2019, pp. 12–23.

[12] A. G. Regino, R. O. Caus, V. Hochgreb e J. C. dos Reis, "QART: A Framework to Transform Natural Language Questions and Answers into RDF Triples.," em *KDIR*, 2022, pp. 55–65.

[13] D. Juric, G. Stoilos, A. Melo, J. Moore e M. Khodadadi, "A system for medical information extraction and verification from unstructured text," em *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 13 314–13 319.