UNIVERSIDADE FEDERAL DE GOIÁS (UFG)
INSTITUTO DE INFORMÁTICA (INF)
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO (PPGCC)

RICARDO PEREIRA DE SOUZA FILHO

Reconhecimento de Entidades Nomeadas em Editais de Licitação

GOIÂNIA
2024

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

1. Identificação do material bibliográfico

[X] Dissertação [ ] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

Exemplos: Estudo de caso ou Revisão sistemática ou outros formatos.

2. Nome completo do autor

Ricardo Pereira de Souza Filho

3. Título do trabalho

Reconhecimento de Entidades Nomeadas em Editais de Licitação

4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)

Concorda com a liberação total do documento [X] SIM [ ] NÃO[1]

[1] Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
a) consulta ao(à) autor(a) e ao(à) orientador(a);
b) novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.

Documento assinado eletronicamente por **Nadia Felix Felipe Da Silva**, **Professora do Magistério Superior**, em 28/12/2024, às 16:25, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ricardo Pereira De Souza Filho**, **Discente**, em 04/01/2025, às 09:52, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5069407** e o código CRC **73EB46E9**.

RICARDO PEREIRA DE SOUZA FILHO

Reconhecimento de Entidades Nomeadas em Editais de Licitação

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Informática da Universidade Federal de Goiás, como requisito para obtenção do título de Mestre em Ciência da Computação.
Área de concentração: Ciência da Computação.
Linha de pesquisa: Sistemas Inteligentes e Aplicações.

Orientadora: Profa. Dra. Nádia Félix Felipe da Silva

GOIÂNIA
2024

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Souza Filho, Ricardo Pereira de
Reconhecimento de Entidades Nomeadas em Editais de Licitação [manuscrito] / Ricardo Pereira de Souza Filho. - 2024.
lxii, 62 f.

Orientador: Profa. Dra. Nádia Félix Felipe da Silva.
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2024.
Apêndice.
Inclui siglas, lista de figuras, lista de tabelas.

1. Processamento de Linguagem Natural. 2. Reconhecimento de Entidades Nomeadas. 3. Editais de Licitação. I. Silva, Nádia Félix Felipe da, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

ATA DE DEFESA DE DISSERTAÇÃO

Ata nº **42** da sessão de Defesa de Dissertação de **Ricardo Pereira de Souza Filho**, que confere o título de Mestre em Ciência da Computação, na área de concentração em Ciência da Computação.

Aos vinte e nove dias do mês de novembro de dois mil e vinte e quatro, a partir das catorze horas, via webconferência, realizou-se a sessão pública de Defesa de Dissertação intitulada **"Reconhecimento de Entidades Nomeadas em Editais de Licitação"**. Os trabalhos foram instalados pela Orientadora, Professora Doutora Nádia Félix Felipe da Silva (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professora Doutora Deborah Silva Alves Fernandes (INF/UFG), membra titular interna; Professora Doutora Ellen Polliana Ramos Souza (UFRPE), membra titular externa. A realização da banca ocorreu por meio de videoconferência. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pela Professora Doutora Nádia Félix Felipe da Silva, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos vinte e nove dias do mês de novembro de dois mil e vinte e quatro.
TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Nadia Felix Felipe Da Silva**, **Professor do Magistério Superior**, em 29/11/2024, às 15:41, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Deborah Silva Alves Fernandes**, **Professora do Magistério Superior**, em 29/11/2024, às 15:41, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Ricardo Pereira De Souza Filho**, **Discente**, em 29/11/2024, às 15:54, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **ELLEN POLLIANA RAMOS SOUZA**, **Usuário Externo**, em 29/11/2024, às 16:56, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4991819** e o código CRC **7564C2B6**.

Referência: Processo nº 23070.051952/2024-49 SEI nº 4991819

Resumo

Filho, Ricardo P. S.. **Reconhecimento de Entidades Nomeadas em Editais de Licitação**. Goiânia, 2024. 63p. Dissertação de Mestrado. Programa de Pós-Graduação em Ciência da Computação(PPGCC), Instituto de Informática (INF), Universidade Federal de Goiás (UFG).

Este trabalho explora o uso de modelos de linguagem natural (LLMs) para extração de informações em editais de licitação, focando na tarefa de Reconhecimento de Entidades Nomeadas (REN). Dada a natureza diversa e não padronizada dos editais, o trabalho propõe uma metodologia que integra técnicas de seleção semântica e cenários de Zero-Shot e Few-Shot, com o objetivo de otimizar o processo de anotação e extração de entidades, reduzindo a necessidade de intervenção manual e melhorando a precisão.

O primeiro passo foi a construção de um *corpus* anotado com entidades nomeadas em editais de licitação. Em seguida, os modelos BERTimbau, BERTikal e mDeBERTa foram treinados supervisionadamente neste conjunto de dados anotado. Os experimentos mostraram que o BERTimbau apresentou melhor desempenho geral, alcançando valores acima de 0.80 para a métrica de avaliação F1-score. Nos cenários Zero-Shot e Few-Shot, diferentes *templates* de *prompt* e estratégias de seleção de exemplos foram testados. Modelos como GPT-4 e LLaMA obtiveram desempenho equivalente aos modelos que passaram por treinamento supervisionado com o auxílio de exemplos semanticamente relevantes, apesar de resultados modestos no cenário sem exemplos.

Os resultados indicam que a combinação de *prompts* enriquecidos com exemplos e a pré-seleção de sentenças relevantes na etapa de anotação contribui para maior precisão e eficiência do processo de REN em editais de licitação. A metodologia apresentada pode ser aplicada para extração de informações, com potencial impacto na transparência e auditoria de licitações públicas.

Palavras–chave

Processamento de Linguagem Natural, Reconhecimento de Entidades Nomeadas, Editais de Licitação

Abstract

Filho, Ricardo P. S.. **T**. Goiânia, 2024. 63p. MSc. Dissertation. Programa de Pós-Graduação em Ciência da Computação(PPGCC), Instituto de Informática (INF), Universidade Federal de Goiás (UFG).

This work explores the use of large language models (LLMs) for information extraction in public procurement notices, focusing on the Named Entity Recognition (NER) task. Given the diverse and unstandardized nature of these documents, the study proposes a methodology that integrates semantic selection techniques with Zero-Shot and Few-Shot scenarios, aiming to optimize the annotation and entity extraction process, reduce manual intervention, and improve accuracy.

The first step involved building an annotated corpus containing named entities from procurement notices. Subsequently, the BERTimbau, BERTikal, and mDeBERTa models were trained in a supervised manner using this annotated dataset. Experiments showed that BERTimbau achieved the best overall performance, with an F1-score above 0.80. In the Zero-Shot and Few-Shot scenarios, various prompt templates and example selection strategies were tested. Models such as GPT-4 and LLaMA achieved performance comparable to supervised models when aided by semantically relevant examples, despite modest results in the absence of examples.

The results indicate that combining enriched prompts with examples and the pre-selection of relevant sentences during the annotation phase contributes to greater accuracy and efficiency in the NER process for procurement notices. The proposed methodology can be applied to information extraction, with potential impacts on transparency and auditing in public procurement.

Keywords

Natural Language Processing, Named Entity Recognition, Procurement Notices

Lista de Siglas e Abreviaturas

BERT *Bidirectional Encoder Representations from Transformers*

BiLSTM *bidirectional Long Short-Term Memory*

CBoW *Continuous Bag-of-Words*

CNN *Convolutional Neural Networks*

CRF *Conditional Random Fields*

LLM *Large Language Model*

LSTM *Long Short-Term Memory*

MLP *Multilayer Perceptron*

PLN Processamento de Linguagem Natural

REN Reconhecimento de Entidades Nomeadas

RNA Redes Neurais Artificiais

RNN *Recurrent Neural Networks*

RNP Redes Neurais Profundas

TCMGO Tribunal de Contas dos Municípios de Goiás

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

No domínio jurídico, a quantidade massiva de coleções de documentos torna desafiadora a análise manual dessas demandas. Nos órgãos fiscalizatórios do governo, por exemplo, os processos de auditoria envolvem a extração de informações em centenas de documentos para averiguar se os órgãos da Administração Pública seguem os princípios da legalidade, da legitimidade e da economicidade na execução orçamentária e financeira dos recursos públicos. Considerando que a extração dessas informações requer compreensão de texto e conhecimentos específicos sobre o domínio desses documentos, automatizar essa tarefa exige sistemas inteligentes capazes de aprender padrões no texto que permitam a correta identificação dessas informações.

Nesse sentido, o Processamento de Linguagem Natural (PLN) é o subcampo da Inteligência Artificial (IA) que possibilita aos computadores executar tarefas que exijam o entendimento e a manipulação de linguagem natural, escrita ou falada [14]. Dentre as tarefas de PLN para Extração de Informação (EI), a tarefa de Reconhecimento de Entidades Nomeadas (REN) parece ser a mais adequada ao problema. A tarefa de REN tem por objetivo reconhecer no texto palavras ou frases que identifiquem um item, ou Entidade Nomeada (EN)[89]. Cada EN pertence necessariamente a uma classe, categoria ou tipo. As entidades podem trazer consigo informações importantes sobre o conteúdo de documentos e, portanto, servir de insumos para outras tarefas, como Recuperação de Informação e Classificação de Documentos, ou mesmo auxiliar na tomada de decisão.

Nesse trabalho, em parceria com o Tribunal de Contas dos Municípios do Estado de Goiás (TCMGO), avaliamos a utilização de sistemas inteligentes para automatizar a extração de informações de editais de licitação. Primeiramente, adotou-se uma abordagem baseada em Aprendizado de Máquina, em que modelos da família BERT[21] foram treinados para a tarefa de REN a partir de um *corpus* composto por textos de editais de licitação em que os dados de interesse foram anotados manualmente como entidades nomeadas. Em uma segunda abordagem, os modelos de linguagem generativos foram avaliados na extração dos dados em cenários com restrição de dados anotados.

1.1 Motivação

No Brasil, o Tribunal de Contas é o órgão de natureza administrativa responsável pela fiscalização do emprego de recursos públicos [39]. Uma das funções do Tribunal é auditar os editais de licitações, nos quais a Administração descreve as regras do procedimento que selecionará a melhor proposta visando a contratação de bens e serviços [20]. Em nível estadual, o Tribunal de Contas é responsável por processar os editais de cada município pertencentes ao seu respectivo ente federativo, dentro de um prazo estabelecido por lei.

Considerando apenas o trabalho desempenhado pela Secretaria de Fiscalização de Engenharia (SFE) do TCMGO, foram avaliados 154 editais de licitação de obras e serviços de engenharia no ano de 2023, resultando em correções em vários deles, incluindo alterações nos valores de 15 editais, resultando em uma economia potencial de 34,1 milhões de reais para os municípios. Segundo o tribunal, "as principais irregularidades detectadas foram incompatibilidades entre modalidade, preço e prazo de publicação, condições inadequadas no projeto básico, e cláusulas restritivas".

Com a digitalização do processo de submissão de editais, os municípios fornecem, além dos editais digitalizados, as principais informações da licitação à parte, preenchidas em formulário. Ainda assim, é preciso conferir se os dados repassados estão de acordo com o conteúdo do documento, além de realizar as demais validações referentes ao trabalho de auditoria.

Nesse cenário, um sistema capaz de extrair, com confiabilidade, as principais informações das licitações a partir dos textos de editais poderia ser utilizado tanto para facilitar o trabalho dos auditores quanto para validar automaticamente os dados repassados pelos usuários do sistema em que os editais são submetidos ao Tribunal de Contas. Além disso, facilitaria a identificação de casos em que há incompatibilidades entre modalidade, preço e prazo de publicação, podendo assim aumentar consideravelmente o número de processos analisados, como também a eficiência da fiscalização do tribunal.

Outro aspecto desafiador é a incorporação de novas informações de interesse à solução, ou seja, novos tipos de entidade, sem que isso exija necessariamente a anotação de um grande volume de dados. Isso se dá porque o convênio com o Tribunal prevê que haja transferência de tecnologia para a equipe de Tecnologia da Informação do Tribunal, de modo que o próprio Tribunal seja capaz de expandir a solução para outros tipos de dados que possam ser extraídos dos editais de licitação.

1.2 Objetivo

O objetivo do projeto é, portanto, automatizar a extração das principais informações do processo licitatório contidas nos editais de licitação, considerando cenários com diferente disponibilidade de dados anotados. Como objetivos secundários, destacam-se:

- Construção de um *corpus* composto por textos extraídos de editais e anotado com as entidades de interesse para treinamento e avaliação de modelos de inteligência artificial.
- Comparar o desempenho de modelos treinados no *corpus* anotado com o desempenho de LLMs generativas em cenários com pouco ou nenhum dado anotado.

1.3 Questões de Pesquisa

Considerando o problema apresentado, esse trabalho se propõe a responder às seguintes questões de pesquisa:

Questão 1: Havendo um *corpus* de editais de licitação com as informações de interesse anotadas como entidades nomeadas, é viável treinar modelos de Reconhecimento de Entidades Nomeadas com desempenho satisfatório para extração de tais informações?

Questão 2: Considerando cenários em que não há grande disponibilidade de dados, os modelos de linguagem generativos podem representar uma alternativa para extração de informações com desempenho similar aos modelos treinados com dados anotados?

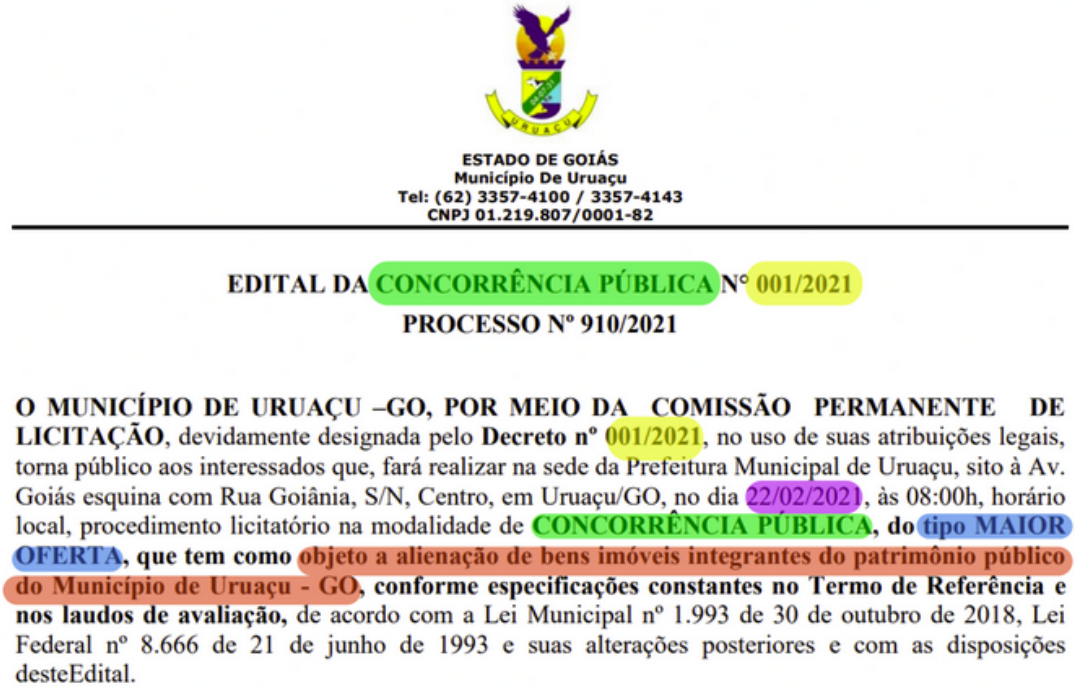
Questão 3: Qual o impacto de diferentes *templates* de *prompt* e estratégias de exemplo no desempenho dos modelos em cenários *Zero-Shot* (nenhum exemplo) e *Few-Shot* (poucos exemplos) ?

Questão 4: Quais classes de entidade apresentam maior dificuldade para extração automática e por quê?

1.4 O domínio

O domínio é constituído por textos de editais de licitações nos quais o Poder Executivo detalha o processo de aquisição de bens e serviços por meio de licitação. As informações básicas sobre o processo licitatório incluem: a modalidade da licitação, o objeto da licitação, o valor do objeto, o número/exercício (número identificador do edital), critério de julgamento e data da sessão. A *modalidade de licitação* **MOD** define as regras do procedimento licitatório, cada uma com suas próprias particularidades. Segundo a

Lei 14.133/2021, existem cinco tipos de modalidades: concorrência, concurso, diálogo competitivo, leilão e pregão. O *objeto da licitação* **OBJ** é o bem ou serviço que se deseja adquirir por meio da licitação e o *valor do objeto* **VLR** é o gasto estimado na aquisição do objeto. O *número/exercício* **NUM** é o código identificador da licitação formado por um número sequencial e pelo ano de exercício de lançamento do edital. O *critério de julgamento* **CRT** é o critério utilizado para escolha da melhor proposta. E por fim, a *data de sessão* **DT** informa o dia em que o processo licitatório ocorrerá.

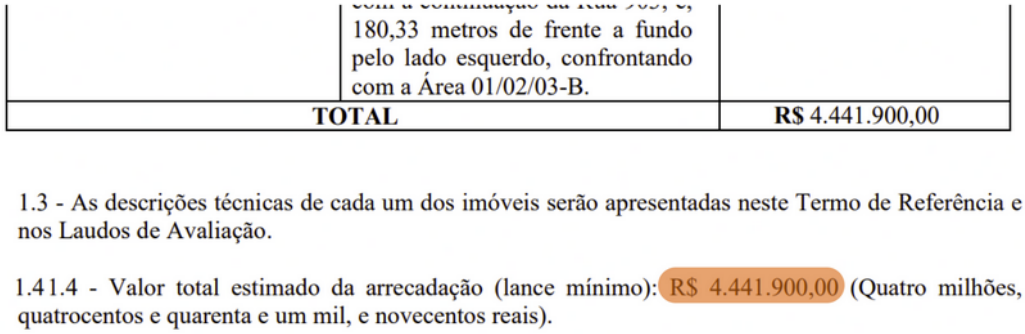


ESTADO DE GOIÁS
Município De Uruaçu
Tel: (62) 3357-4100 / 3357-4143
CNPJ 01.219.807/0001-82

EDITAL DA CONCORRÊNCIA PÚBLICA Nº 001/2021
PROCESSO Nº 910/2021

O MUNICÍPIO DE URUAÇU –GO, POR MEIO DA COMISSÃO PERMANENTE DE LICITAÇÃO, devidamente designada pelo **Decreto nº** 001/2021, no uso de suas atribuições legais, torna público aos interessados que, fará realizar na sede da Prefeitura Municipal de Uruaçu, sito à Av. Goiás esquina com Rua Goiânia, S/N, Centro, em Uruaçu/GO, no dia 22/02/2021, às 08:00h, horário local, procedimento licitatório na modalidade de **CONCORRÊNCIA PÚBLICA, do tipo MAIOR OFERTA, que tem como objeto a alienação de bens imóveis integrantes do patrimônio público do Município de Uruaçu - GO, conforme especificações constantes no Termo de Referência e nos laudos de avaliação,** de acordo com a Lei Municipal nº 1.993 de 30 de outubro de 2018, Lei Federal nº 8.666 de 21 de junho de 1993 e suas alterações posteriores e com as disposições desteEdital.

Figura 1.1: *Exemplo de entidades presentes no preâmbulo do edital de licitação: NUM (amarelo), DT (lilás), CRT (azul) e OBJ (laranja).*



	180,33 metros de frente a fundo pelo lado esquerdo, confrontando com a Área 01/02/03-B.	
TOTAL		**R$** 4.441.900,00

1.3 - As descrições técnicas de cada um dos imóveis serão apresentadas neste Termo de Referência e nos Laudos de Avaliação.

1.41.4 - Valor total estimado da arrecadação (lance mínimo): R$ 4.441.900,00 (Quatro milhões, quatrocentos e quarenta e um mil, e novecentos reais).

Figura 1.2: *Exemplo de valor do objeto em texto e tabela.*

Não há um padrão rígido a ser seguido para escrita e formatação dos editais, assim os documentos submetidos ao Tribunal de Contas contêm uma diversidade de ca-

beçalhos, rodapés, assinaturas eletrônicas, padrões de numeração de sessões, imagens, tabelas e anexos. Há também documentos físicos, posteriormente digitalizados, e portanto, necessitam de tecnologias de imagem para extração do texto. Apesar disso, em geral, as informações principais se encontram no início do texto (capa e preâmbulo) (Figura 1.1), com a exceção do valor do objeto que pode ter uma seção específica no decorrer do edital (Figura 1.2).

1.5 Organização da Dissertação

O restante desse trabalho está organizado conforme a seguinte estrutura: o Capítulo 2 apresenta a fundamentação teórica; no Capítulo 3 apresentam-se os trabalhos relacionados; o Capítulo 4 descreve a metodologia utilizada; os experimentos realizados e os resultados obtidos são relatados no Capítulo 5, seguido das conclusões apresentadas no Capítulo 6.

CAPÍTULO 2

Fundamentação Teórica

Nesse capítulo, a tarefa de Reconhecimento de Entidades Nomeadas (REN) é apresentada em detalhes, com ênfase nas abordagens baseadas em *Deep Learning*.

2.1 Reconhecimento de Entidades Nomeadas

Reconhecimento de Entidades Nomeadas (REN) é uma tarefa de Processamento de Linguagem Natural cujo objetivo é identificar menções no texto que pertencem a tipos semânticos [62] . No primeiro trabalho a mencionar o conceito de Entidade Nomeada (EN) [31], por exemplo, foram considerados os seguintes tipos de entidades: organização, pessoa e local geográfico. De maneira formal, dada uma sentença s, composta por N *tokens* $s = <w_1, w_2, w_3, ..., w_N>$, o objetivo da tarefa é obter a lista de entidades presentes no texto representada pelas tuplas $<I_s, I_e, t>$, em que $I_s \in [1, N]$ e $I_e \in [1, N]$ são respectivamente os índices de início e fim da entidade, e t é o tipo da entidade.

Quanto à definição de Entidade Nomeada (EN), os autores divergem quanto a rigidez do conceito, limitando-o aos substantivos próprios ou incluindo outros tipos de menções às entidades [41, 62, 69]. No entanto, de maneira geral, a expressão pode ser definida como "um substantivo próprio, que serve como nome de algo ou alguém"[69]. Em muitos trabalhos, esse conceito é extrapolado, as ENs não se limitam a substantivos próprios, mas também são utilizados para identificar outros elementos de texto como datas e valores monetários, por exemplo. Portanto, de maneira mais ampla, a Entidade Nomeada é a palavra ou frase que identifica um item ou um conjunto de itens que pertencem a um determinado tipo semântico ou que têm atributos em comum [89].

Já quanto aos tipo de ENs, existem os tipos genéricos ou universais: pessoa, local, organização; e os tipos de domínio específico, como por exemplo, no domínio legal: juiz, legislação, valor de condenação, tribunal.

2.1.1 Métricas de Avaliação

A avaliação dos sistemas REN é feita comparando suas saídas a dados anotados manualmente. A comparação é quantificada ao calcular os números de Falsos Positivos (FP), Falsos Negativos (FN) e Positivos Verdadeiros (PV).

- Falsos Positivos são entidades identificadas pelo sistema REN que não aparecem nos dados anotados.
- Falsos Negativos são entidades presentes nos dados anotados manualmente, mas que não foram identificadas pelo sistema REN.
- Positivos Verdadeiros são entidades identificadas pelo sistema REN e que também constam nos dados anotados.

As quantidades de Falsos Positivos, Falsos Negativos e Positivos Verdadeiros são utilizadas para calcular as métricas de avaliação, Precisão, Recall e F-score:

$$Precision = \frac{PV}{PV+FP}, Recall = \frac{PV}{PV+FN}, Fscore = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- Precisão é o percentual de entidades corretamente identificadas dentre as que foram retornadas pelo sistema REN, ou seja, é a proporção de Verdadeiros Positivos identificados pelo sistema que de fato são verdadeiros.
- Recall é o percentual de entidades corretamente identificadas dentre o total de entidades presentes no dados anotados.
- F-score é a média harmônica entre Precisão e Recall. A métrica de F-score pode ser calculada pela média de F-score para cada tipo de entidade(*Macro-Averaged F-score*); ou considerando os números FP, FN e PV entre todos os tipos de entidades (*Micro-Averaged F-score*). No caso de Micro-Averaged F-score, os resultados são mais sensíveis à performance do sistema REN nos tipos de entidades com maior número de ocorrências no conjunto de testes.

A tarefa de REN pode ser dividida em duas sub-tarefas: a detecção dos limites das entidades e a identificação do tipo da entidade. Portanto, há dois tipos de avaliação: *Exact-match Evaluation*, em que as entidades são contabilizadas como corretas se tanto o tipo como os limites das entidades forem exatos; e *Relaxed-match Evaluation*[31], em que a entidade é considerada correta se o tipo estiver correto e houver ao menos uma sobreposição entre os limites identificados pelo sistema REN e os limites indicados pela anotação manual.

2.1.2 Corpora para REN em Português

A Tabela 2.1 apresenta os estudos nos quais um novo corpus para REN em Português foi criado ou atualizado até Junho de 2022 [4], os trabalhos estão classificados

quanto ao método de anotação utilizado, o domínio ao qual os textos do *corpus* pertencem, bem como a sua variante do Português.

As informações demonstram um aumento no interesse da tarefa ao longo do tempo, bem como a diversificação dos domínios abordados. Dentre os domínios específicos, o domínio legal/legislativo tem o maior *corpora*. Um olhar revela a diversidade de subgêneros dentro desse domínio: documentos produzidos pela Suprema Corte brasileira [51], leis [51], portarias [82], propostas de lei e solicitações de trabalho da Câmara dos Deputados do Brasil [3].

Corpus	**Ano**	**Número de tipos**	**Método de anotação**	**Domínio**	**Variante da Língua**
Second HAREM[28]	2010	10	Manual	Geral	PT-BR & PT-EU
[58]	2011	3	Automático	Journalistic	PT-EU
Summ-it++[27]	2016	10	Automático	Geral	PT-BR
GeoCorpus[8]	2017	13	Manual	Geologia	PT-BR
Dicionário Histórico-Biográfico Brasileiro (DHBB)[36]	2018	7	Híbrido	História	PT-BR
LeNER-Br[51]	2018	6	Manual	Legal	PT-BR
SESAME[54]	2019	3	Automático	Geral	PT-BR
[49]	2019	13	Manual	Médico	PT-EU
[60]	2019	3	Manual	Polícia	PT-BR
[97]	2019	2	Híbrido	Informações de Tráfego	PT-BR
[83]	2019	2	Manual	Literatura	PT-BR
DataSense NER Corpus[22]	2020	18	Manual	Dados Sensíveis	PT-EU
DrugSeizures-Br[10]	2020	6	Automático	Legal	PT-BR
GeoCorpus-2[17]	2020	13	Manual	Geologia	PT-BR
[19]	2020	3	*Não Informado*	Jornalístico	PT-BR
[82]	2020	1	Manual	Legal	PT-BR
[29]	2020	4	Manual	Geral	PT-EU
Aposentadoria[65]	2021	10	Híbrido	Legal	PT-BR
EHR-Names[86]	2021	1	Manual	Médico	PT-BR
Financial Market Corpus [81]	2021	3	Híbrido	Financeiro	PT-BR
[9]	2021	6	Híbrido	Legal	PT-BR
[61]	2021	2	Manual	Legal	PT-BR
[91]	2021	10	Manual	E-commerce	PT-BR
UlyssesNER-Br[3]	2022	7	Manual	Legislativo	PT-BR

Tabela 2.1: *Lista do corpora de REN em Português. Adaptado de [4]*

2.1.3 Abordagens tradicionais de REN

Abordagens de REN baseadas em regras

A abordagem de REN baseada em regras faz uso de regras desenvolvidas manualmente para identificação das entidades. Parte das regras se baseiam na detecção de padrões de sintáticos ou semânticos que indicam a presença das entidades no texto

[79, 15]. Outra estratégia utilizada em contextos restritos a um domínio específico, é a utilização de dicionários de termos específicos ao domínio (*gazetteers*) [25, 87].

Apesar dos sistemas de REN baseados em regras poderem atingir alta precisão em identificar entidades em contextos específicos, em geral, eles têm baixa performance ao serem transferidos para outros domínios, pois são limitados pela capacidade limitada de generalização de suas regras e pela cobertura limitada de seus dicionários.

Técnicas de REN de aprendizado supervisionado

Em abordagens de aprendizado supervisionado, a tarefa de REN é entendida como uma tarefa de classificação ou rotulação multi-classe de sequências. Algoritmos de aprendizado de máquinas são treinados com uso de dados anotados. Cada exemplo dos dados anotados possui características, ou *features*, projetadas para representar o texto de modo que o algoritmo tenha informações relevantes para aprender a executar a tarefa.

A escolha do melhor conjunto de *features* é fundamental para o sucesso do aprendizado supervisionado. Cada palavra que compõe a sequência de entrada é representada por um vetor composto por valores booleanos, numéricos ou nominais [62]. Exemplos comuns de *features* são: capitalização, classe gramatical, informações morfológicas, pontuação e número de ocorrências da palavra no documento [107, 88]. Informações provenientes de fontes externas como *gazeteers* também podem ser incorporadas às *features*[73, 80].

Quanto aos algoritmos de aprendizado supervisionado, diferentes métodos foram aplicados a REN, como: *Hidden Markov Models* (HMM) [24], Árvores de Decisão [76], Support Vector Machines (SVM) [35], e Conditional Random Fields(CRF) [43].

2.2 Técnicas de *Deep Learning* para REN

Aprendizado Profundo ou *Deep Learning*, é uma área de Aprendizado de Máquina que estuda métodos que utilizam múltiplas camadas de representação de dados [30]. Cada camada é uma rede neural que produz uma abstração da camada anterior, começando a partir dos dados brutos. As múltiplas transformações produzidas pelas camadas resultam em um mapeamento não-linear dos dados, possibilitando o aprendizado de funções complexas. Para tarefas de classificação, como REN, camadas superiores de abstração potencializam aspectos latentes aos dados que são relevantes para a classificação e paralelamente suprimir informações irrelevantes. Durante o treinamento supervisionado, as camadas neurais têm seus parâmetros ajustados pelo gradiente descendente de um função objetivo que quantifica o erro obtido nos dados de treinamento[30].

Portanto, os métodos de *Deep Learning* proporcionam as seguintes vantagens frente às abordagens tradicionais:

- Possibilita o aprendizado de funções complexas, via a transformação não-linear das representações dos dados.
- Dispensa a engenharia de *features*, já que os métodos de *Deep Learning* se utilizam dos próprios dados brutos.
- Treinamento de ponta a ponta por meio de gradiente descendente para projetar sistemas REN complexos.

A arquitetura genérica de um sistema de REN baseado em métodos de *Deep Learning* é composta pelos três elementos apresentados na Figura 2.1, são eles: 1) Representação distribuída para entrada, 2) Codificador de contexto e 3) Decodificador de rótulo (Figura 2.1) [45]. Esses elementos são descritos em mais detalhes nas seções seguintes.

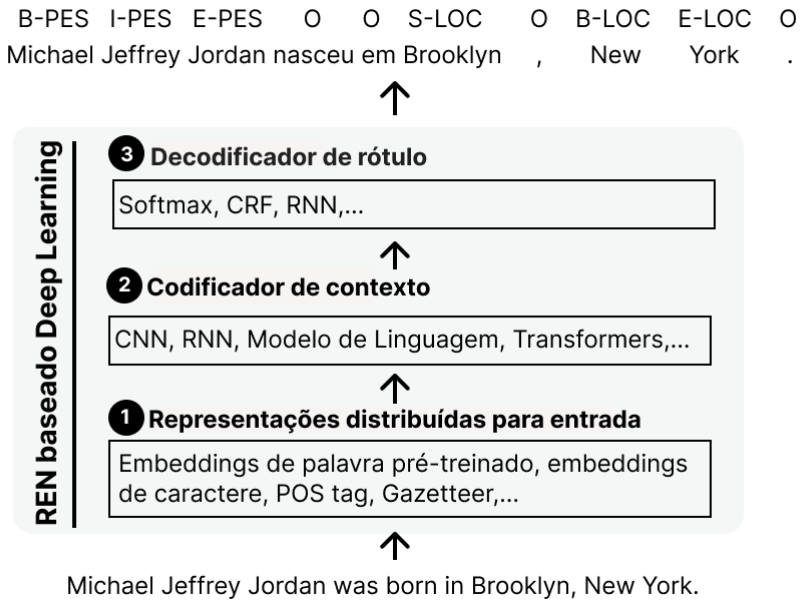


Figura 2.1: *A taxonomia de REN baseado em Deep Learning. Adaptado de [45]*

2.2.1 Representações distribuídas para entrada

As representações distribuídas, também chamadas de *embeddings*, são vetores densos de baixa dimensão usados para representar palavras ou caracteres em que cada dimensão representa uma característica latente do texto representado, como informações sintáticas e semânticas [6]. Essas informações latentes são construídas automaticamente por redes neurais através de treinamento não-supervisionado em um grande corpora, esse processo é chamado de pré-treino. A ideia fundamental desses treinamentos é fazer uso dos contextos em que as palavras se encontram para aprender os seus significados e funções sintáticas [6]. Os *embeddings* podem representar texto a nível de palavra

ou caractere, e ainda serem híbridas, combinando as informações a nível de palavra, caractere, entre outras informações.

Representações vetoriais a nível de palavras

Os *embeddings* de palavras podem ser classificados pelo tipo de algoritmo não-supervisionado utilizado para o pré-treino. Os primeiros estudos utilizaram os algoritmos *continuous bag-of-words* (CBOW) [66] e *continuous skip-gram* [56]. Outros exemplos de embeddings de palavras amplamente utilizados são Word2Vec [57], Glove [68] e fastText [38].

Representações vetoriais a nível de caracteres

Para os sistemas que utilizam *embeddings* de caracteres, as palavras são representadas pela combinação dos *embeddings* que representam cada uma das letras que as compõem. Nesses sistemas, os *embeddings* de caracteres são produzidos por modelos baseados em redes neurais convolucionais [52, 47, 71] ou recorrentes [42, 44] que extraem informações a nível de caractere, depois as representações de caractere obtidas são concatenadas ou comprimidas em um único *embedding* de palavra [52].

Em comparação com os *embeddings* de palavras, os *embeddings* de caracteres são capazes de incorporar mais informações morfológicas das palavras, como sufixos e prefixos e flexões verbais, como também contornam o problema de palavras fora do vocabulário, já que os sistemas de *embeddings* de palavras são limitados por um vocabulário finito, afinal qualquer palavra pode ser representada a partir das representações de suas letras.

Representações vetoriais híbridas

Representações híbridas são aquelas que adicionam informações de diferentes fontes, para além das representações de palavras, como: informações gramaticais, informações obtidas em dicionários, informações morfológicas e maiusculização da primeira letra, entre outras características da palavra [16, 101, 1]. Destaca-se também entre as representações híbridas a proposta em por [21], em que as palavras são dividas em *tokens*, cada um com uma representação vetorial pré-treinada, no entanto, a representação final da palavra também inclui o somatório das informações de posição na frase e segmento.

2.2.2 Arquiteturas de Codificadores de Contexto

Redes Neurais Convolucionais

Proposto por [16], o codificador de contexto baseado em redes neurais convolucionais (Figura 2.2) considera toda a sentença para determinar o rótulo de uma palavra. Cada palavra é representada por um vetor de N dimensões. Então, a camada convolucional é responsável por extrair características de cada palavra considerando o contexto da sentença, gerando como saída uma representação local e intermediária de cada palavra. Depois, uma representação global é construída pela combinação das representações locais. Dentre as operações mais comumente utilizadas para produzir a representação global, estão a média e *max* ao longo das posições das palavras na sentença. Por fim, as representações globais são utilizadas por um decodificador de rótulos 2.2.3 para determinar os rótulos de cada palavra. Redes convolucionais também são usadas em soluções híbridas a fim de atenuar problemas das redes recorrentes, como memória de curto prazo [108] e eficiência computacional [95].

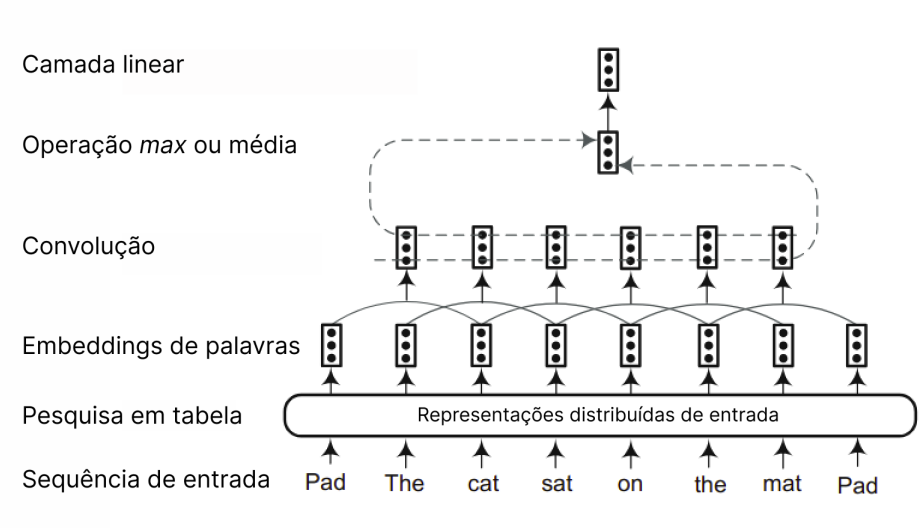


Figura 2.2: *Rede neural convolucional aplicada a uma sentença. As características extraídas para cada palavra consideram a totalidade da sentença. Adaptado de [45]*

Redes Neurais Recorrentes

A arquitetura das redes neurais recorrentes têm sido amplamente utilizada para aplicações com dados sequenciais: séries temporais e texto. Especialmente as versões bidirecionais (Figura 2.3), bidirecional long-short term memory (BiLSTM), foram bastante exploradas para tarefas de PLN de classificação de sequência, como REN. As BiLSTM oferecem a vantagem de considerar a informação anterior como também a informação posterior à palavra ao classificá-la.

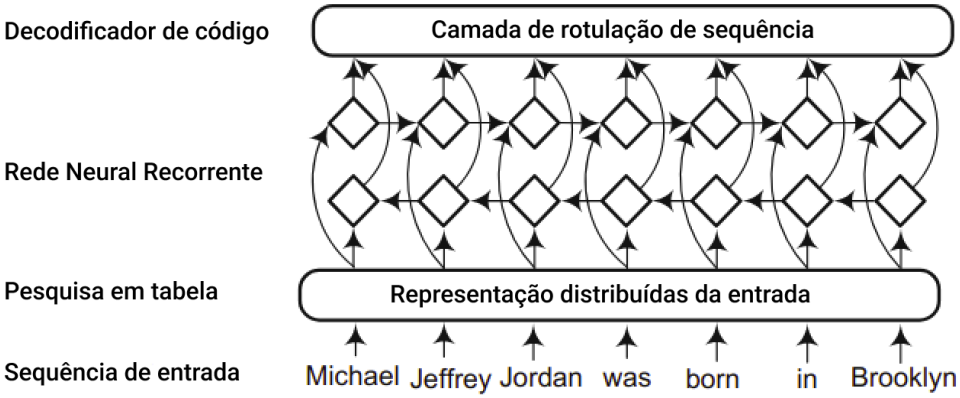


Figura 2.3: *Arquitetura de um codificador de contexto baseado em redes neurais recorrentes. Adaptado de [45]*

Arquitetura *Transformers*

A arquitetura de redes neurais denominada *Transformers* [98] foi proposta inicialmente para tarefa de tradução automática. Os modelos de tradução automática, em geral, apresentam estrutura codificador-decodificador (Figura 2.4) em que o módulo codificador toma como entrada um sequência de *tokens* $(x_1, \ldots, x_n)$ e a mapeia em uma sequência de representações $z = (z_1, ..., z_n)$, a partir de z o decodificador gera a sequência de *tokens* de saída $y = (y_1, \ldots, y_m)$ já traduzida, um elemento de cada vez. Até então, os modelos neurais mais bem sucedidos para tradução se baseavam em redes recorrentes [96], ou convolucionais [40]. A arquitetura *Transformer*, no entanto, se baseia em outro mecanismo neural denominado **Atenção**, proposto por [13] e [50].

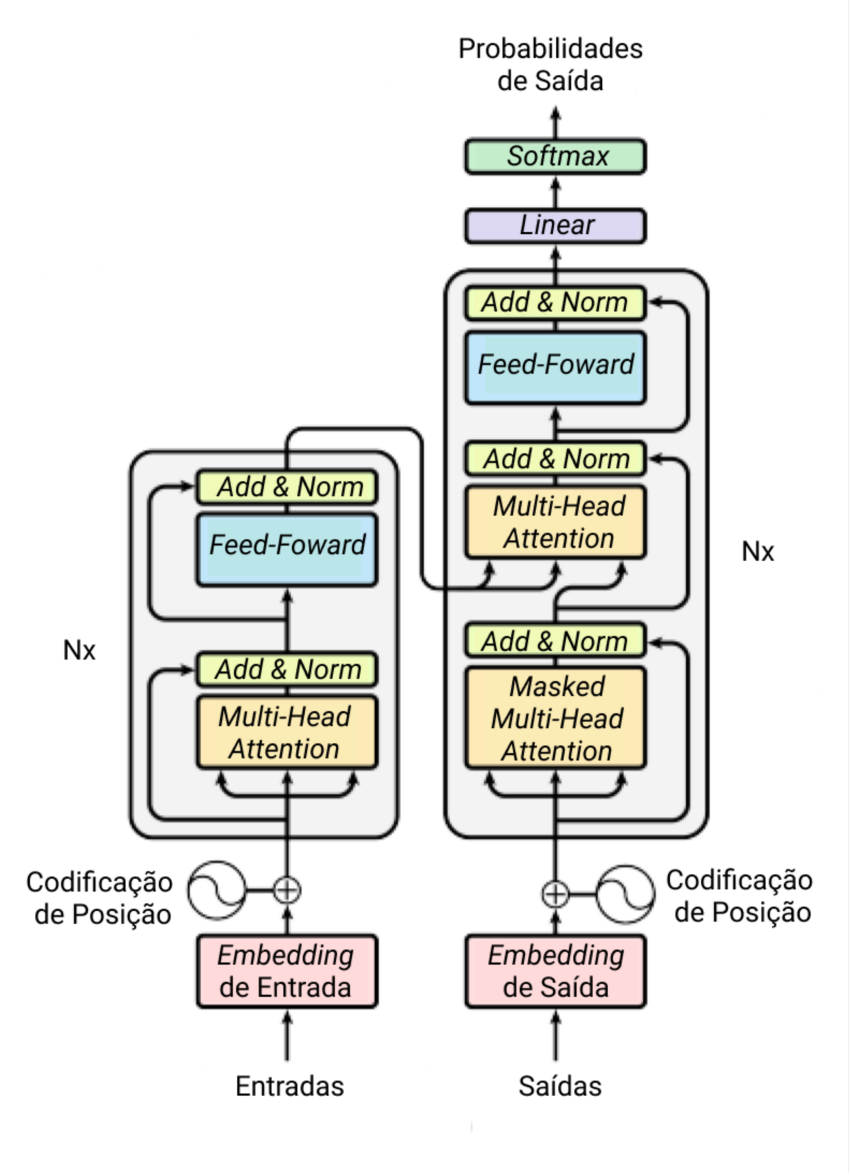


Figura 2.4: *Arquitetura do modelo Transformers, adaptado e traduzido de [98]*

Na arquitetura *Transformer* (Figura 2.4), codificador e decodificador são formados por seis camadas idênticas. Cada camada se subdivide em duas subcamadas: uma subcamada de *Multi-Head Attention* com oito funções de atenção em paralelo, e a outra é uma subcamada totalmente conectada *feed-forward*. As saídas de cada subcamada são normalizadas e uma ligação residual conecta diretamente a entrada da subcamada à camada de normalização. As saídas das subcamadas bem como os *embeddings* de entrada e saída do modelo possuem a mesma dimensão $d_{model} = 512$.

Dentre os modelos de linguagem baseados na arquitetura *Transformers* destacam-se: *Generative Pre-trained Transformer* (GPT) [78], modelo que produz linguagem natural em texto, pré-treinado com enfoque nas tarefas de *language understanding*, como tradução automática (*Machine Translation*) e resposta a pergunta (*Question Answering*);

Bidirectional Encoder Representations from Transformers (BERT), modelo baseado no módulo codificador da arquitetura *Transformers* que usa o contexto em ambas direções, em oposição ao GPT que utiliza apenas o contexto da esquerda para direita [21].

Uma vez que os modelos *Transformers* foram pré-treinados em um corpora massivo, os conhecimentos gerais sobre linguagem adquiridos podem ser aproveitados nas mais diversas tarefas de PLN por meio de técnicas de transferência de aprendizado, diminuindo a necessidade de dados anotados específicos para a tarefa alvo. Dentre os tipos de transferência de conhecimento mais comuns estão a adaptação de domínio, aprendizado multilinguagem e aprendizado multitarefa [75].

2.2.3 Decodificador de rótulos

O decodificador de rótulos é o último módulo do sistema de REN, seu objetivo é determinar a sequência de rótulos correspondente à sentença de entrada, a partir das representações contextualizadas produzidas pelo codificador de contexto. Dentre as arquiteturas para decodificadores mais utilizadas estão: Multi-Layer Perceptron (MLP) + Softmax [47, 95], Conditional Random Fields (CRFs) [37, 106], Redes Neurais recorrentes (RNNs) [90, 66] e Pointer Networks [99, 46].

CAPÍTULO 3

Trabalhos Relacionados

Em termos do domínio dos textos utilizados, este projeto se aproxima dos demais trabalhos de REN no domínio legal em Português. No entanto, os editais de licitações se enquadram na área do Direito Administrativo, enquanto os trabalhos aqui relacionados utilizam documentos pertencentes a outras áreas do Direito, como Direito Legislativo [51, 3] e Direito Trabalhista [11].

3.1 Reconhecimento de Entidades Nomeadas em Língua Portuguesa

O primeiro trabalho que fez uso de tecnologias de aprendizado de máquina para REN utilizou atributos criados e selecionados manualmente para representar o texto, e um modelo Conditional Random Fields (CRF) como classificador [7]. A essa abordagem, Pirovani e Oliveira [73] adicionaram o uso de regras criadas manualmente. As regras consideram características léxicas e sintáticas para a identificação das entidades, complementando os resultados obtidos pelo modelo CRF.

O trabalho de Santos e Guimarães [84] é o primeiro a propor um modelo de aprendizado profundo para REN. O sistema proposto, chamado CharWNN, amplia a arquitetura proposta em [16] com uma rede convolucional para gerar representações vetoriais (embeddings) a nível de caracteres, que combinadas a representações vetoriais pré-treinadas a nível de palavras compõem os atributos utilizados para classificação das palavras.

Depois, outros trabalhos seguiram abordagens combinando *embeddings* pré-treinados de palavras e modelos convolucionais LSTM (Long Short-Term memory). Santos et al. [86] combina Flair Embeddings [2] e *embeddings* de caracteres gerados por um modelo de linguagem bidirecional pré-treinado, os *embeddings* concatenados alimentam o classificador BiLSTM-CRF. Já Castro et al. [77] fez uso de *embeddings* ELMo [70], que por sua vez também são resultado da combinação de *embeddings* de caracteres com *embeddings* gerados por ML baseado em BiLSTM.

O trabalho de Souza et al. [92] explora o potencial dos modelos baseados na arquitetura Transformers. Mais especificamente o modelo BERT pré-treinado em Português. BERT codifica *embeddings* para cada um dos tokens da sequência de entrada, esses *embeddings* alimentam a camada de saída que gera uma pontuação para cada rótulo de entidade para cada *token*. Dentre as alternativas testadas em [92], CRF foi a arquitetura usada como camada de saída que obteve os melhores resultados. A Tabela 3.1 apresenta a evolução dos sistemas REN ao longo do tempo com base nos resultados obtidos no HAREM no cenário total em que são consideradas 10 tipos de entidades nomeadas.

Arquitetura	F1
CharWNN [84] (**2015**)	71.23%
LSTM-CRF [77] (**2018**)	76.27%
ELMo [11] (**2019**)	78.04%
BiLSTM-CRF+FlairBBP [86] (**2019**)	82.26%
$\text{BERT}_{\text{LARGE}}$-CRF [92] (**2020**)	83.24%

Tabela 3.1: *Evolução dos modelos REN em Português em domínio geral.*

3.2 Reconhecimento de Entidades Nomeadas no domínio legal

O trabalho de Luz de Araújo et al. [51] foi o primeiro a empregar a tarefa de REN no domínio legal. Os autores compuseram um conjunto de dados com 70 documentos de diversas cortes brasileiras, nos quais foram anotadas entidades dos tipos *Legislação* e *Casos Legais*, além das entidades universais: *Pessoas*, *Locais*, *Tempo* e *Organizações*. O melhor modelo utilizado nos experimentos se baseia na arquitetura BiLSTM+CRF e vetores de palavras Glove pré-treinados em Português, alcançando F1-score de 92,53%.

Alle [5] utilizou 470 publicações do Diário Oficial da União para desenvolver um conjunto de dados com entidades anotadas dos tipos: *Cargo*, *Lei*, *Número*, *Processo* e *Valor Monetário*. O melhor resultado, F1-score de 44.5%, foi obtido utilizando a ferramenta de REN do *framework* Apache OpenNLP.

Castro [11] construiu um conjunto de dados com documentos de processos da Justiça Trabalhista. Foram anotados os seguintes tipos de entidades: *Função*, *Fundamento*, *Tribunal*, *Vara*, *Valor de Acordo*, *Valor da Causa*, *Valor de Condenação* e *Valor de Custas*, além das tipos *Local*, *Organização* e *Pessoa*. O melhor resultado foi obtido por um modelo ELMo pré-treinado em textos jurídicos, com média de F1-score de 93.81%.

Albuquerque et al. [3] apresentou o UlyssesNER-Br, um conjunto de dados para REN composto por 150 projetos de lei e 800 consultas legislativas da Câmara dos

Deputados do Brasil. As entidades anotadas pertencem aos tipos *Fundamento de Lei* e *Produtos de Lei*, além das entidades universais. O melhor resultado com F1-score de 81.04% foi alcançado utilizando modelos CRF. Costa et al. [18] expandiu o UlyssesNER com um conjunto de comentários de cidadãos sobre projetos de lei. Os modelos CRF, BiLSTM+CRF e BERT foram usados nos experimentos. O melhor resultado foi obtido pelo modelo BERT com F1-score de 73.90%.

Trabalhos mais recentes têm explorado diferentes aspectos do pré-treinamento, tokenização e ajuste fino de *Transformers*, visando melhorar o desempenho em tarefas de aplicação, como REN [94], além de propor métodos mais eficientes para adaptar modelos a diferentes domínios textuais [53].

3.2.1 In-Context Learning para REN

Os avanços recentes com Modelos de Linguagem Generativa inspiraram muitos estudos aplicando-os a diversas tarefas de PLN[12]. Uma estratégia que tem se mostrado eficaz para melhorar o desempenho dos LLMs nessas tarefas é o Aprendizado no Contexto (In-Context Learning, ICL) [23], onde a informação de contexto é enriquecida adicionando alguns exemplos ao prompt. Alguns trabalhos utilizaram o Aprendizado no Contexto para REN [103, 32, 104]. Para o REN em português, Nunes [67] explorou o paradigma de ICL no domínio Legislativo utilizando LeNER-Br e UlyssesNER-Br como *corpora* para executar experimentos comparativos envolvendo diferentes estratégias de seleção de exemplos.

CAPÍTULO 4

Método

A fim de executar o objetivo de desenvolver um sistema para extração de informações a partir de editais de licitações, optamos por abordar o problema como uma tarefa de Reconhecimento de Entidades Nomeadas, considerando cada uma das informações de interesse como entidades nomeadas. Portanto, as seguintes etapas foram realizadas durante a execução do trabalho:

1. Desenvolvimento de heurísticas para identificação e extração de entidades.
2. Pré-processamento dos textos para eventuais correções na extração a partir dos documentos em PDF, e segmentação dos textos em sentenças completas.
3. Construção do conjunto de dados anotados.
4. Treinamento de modelos REN.
5. Desenvolvimento de uma API para tornar acessível o sistema de REN aos sistemas de tecnologia da informação do Tribunal de Contas.

4.1 O sistema proposto

Além das etapas listadas acima, parte importante do projeto foi a pesquisa por métodos eficazes de extração de texto de arquivos em PDF, além da extração de entidades e itens de licitação presentes em tabelas. A Figura 4.1 apresenta o fluxograma do sistema proposto com destaque para o objeto desse trabalho, o módulo REN-Texto. A primeira etapa do processo é a parte de extração de textos e tabelas dos documentos em PDF, nessa fase é importante determinar se o arquivo em questão é um documento digital, ou se trata de um documento físico que foi posteriormente digitalizado e convertido em formato PDF, já que para o segundo caso foi desenvolvida uma solução mais robusta aos ruídos de imagem comuns a esses casos. Nessa fase, os textos são extraídos por ferramentas de OCR, ou ferramentas específicas para leitura de arquivos PDF que retornam os trechos que compõem cada página; em geral, esses trechos correspondem a parágrafos ou sentenças. Também são identificadas e extraídas todas as tabelas do documento. As tabelas podem gerar ruídos para a extração de texto, mas seu conteúdo pode conter os itens licitados,

como também podem conter as entidades nomeadas, especialmente datas, como é o caso da data de sessão.

As tabelas extraídas seguem para a extração de conteúdo; nessa etapa, a estrutura de linhas e colunas é preservada para facilitar a identificação do conteúdo das tabelas. As informações são então extraídas à parte por meio de heurísticas desenvolvidas manualmente.

Por sua vez, o texto segmentado em sentenças segue para a fase de pré-processamento, responsável por: (1) tentar recompor frases quebradas no processo de extração de texto, (2) remover cabeçalhos e rodapés, (3) remover caracteres não ASCII e sequências de espaços em branco. Essa etapa ajuda a manter a integridade do texto, preservando o contexto de cada sentença, como também evita que sentenças irrelevantes sejam processadas pela fase de REN.

Depois de pré-processado, são aplicadas no texto as técnicas de REN, baseadas em regras e modelos *Transformers*. Embora as técnicas REN sejam dispares entre si, o retorno de ambas segue o mesmo padrão. Para cada entidade, é retornado seu conteúdo, sua classe, a página em que a entidade foi extraída e parte do trecho ao qual a entidade pertence.

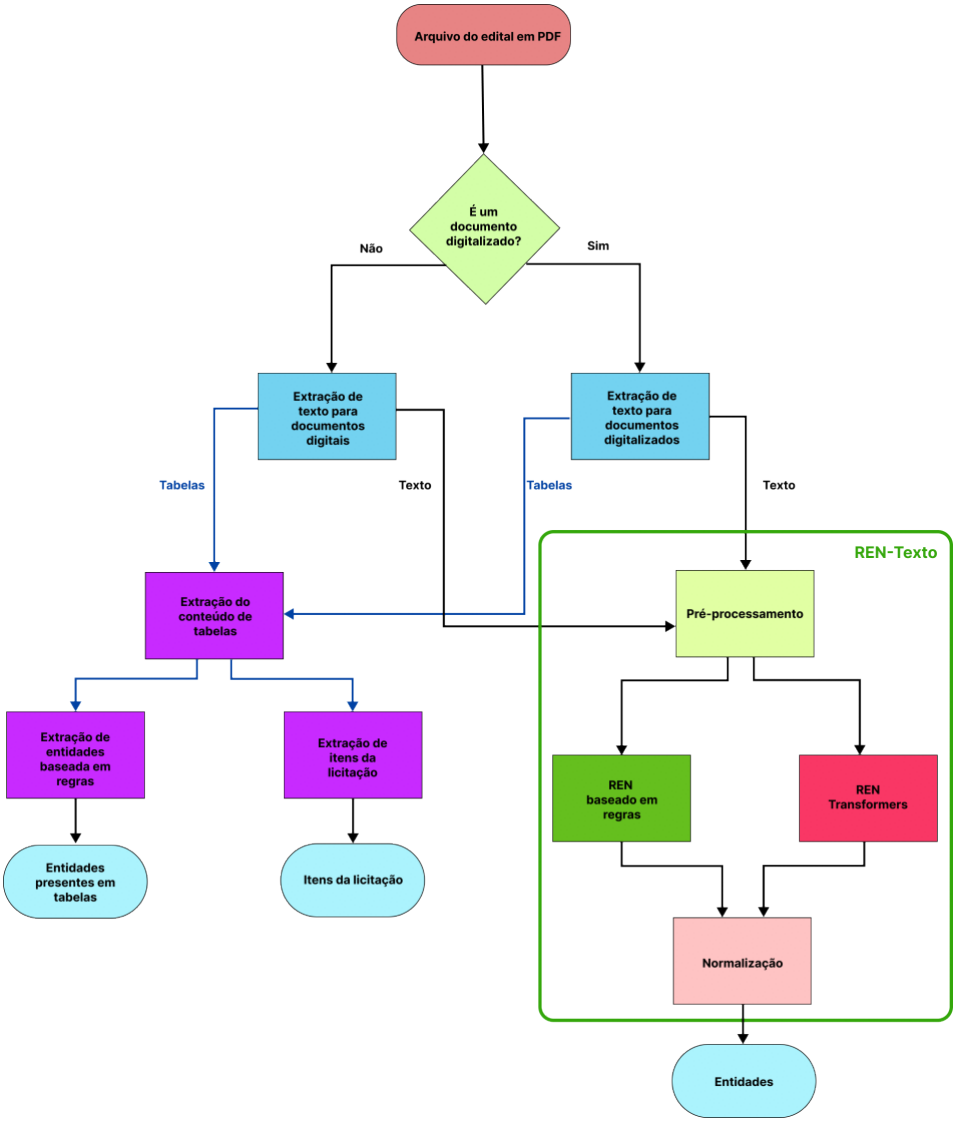


Figura 4.1: *Fluxograma do sistema proposto.*

4.2 Reconhecimento de Entidades Nomeadas baseado em regras

Extrair entidades por meio de regras de extração é, em geral, uma alternativa com limitações de cobertura, especialmente em cenários em que há grande variedade de texto. No entanto, pode ser um ponto de partida enquanto soluções mais sofisticadas estão em desenvolvimento, bem como compor soluções híbridas que combinam as entidades extraídas por múltiplas soluções.

A maioria dos sistemas baseados em regras têm três componentes [59]: (1) um conjunto de regras de extração, (2) um vocabulário, contendo vocábulos de domínio específico, também conhecido como *gazeteers* [48], e (3) uma aplicação das regras e

vocábulos ao texto, como a biblioteca RegEx[1].

4.3 Reconhecimento de Entidades Nomeadas com Modelos Pré-Treinados

Os modelos *Transformers* pré-treinados podem ser armazenados e compartilhados em repositórios *online* como Hugging Face Hub[2]. Os treinamentos supervisionados para diferentes tarefas de PLN são facilitados por *frameworks* específicos para modelos *Transformers*[3].

4.4 Anotação do *Corpus*

O treinamento de modelos de aprendizado profundo requer um conjunto de dados com entidades anotadas. Nesse processo, as entidades nomeadas são identificadas e classificadas entre as classes, que nesse caso correspondem às informações básicas do processo licitatório: modalidade da licitação **MOD**, objeto da licitação **OBJ**, valor do objeto **VLR**, número/exercício **NUM**, critério de julgamento **CRT** e data da sessão **DT**(Figura 4.2).

O MUNICÍPIO DE... torna público que realizará licitação, na modalidade Tomada de Preços **MOD** , autuada sob o nº 006/2021 **NUM** , DO TIPO MENOR PREÇO POR EMPREITADA GLOBAL **CRT**, visando a contração de empresa para execução de cobertura metálica na Unidade de Pronto Atendimento **OBJ** ... tendo seu valor estimado em R$ 24.582,19 (vinte e quatro mil, quinhentos e oitenta e dois reais e dezenove centavos) **VLR** ... DATA DE SESSÃO: 22/03/2020 **DT** ...

Figura 4.2: *Exemplos das entidades anotadas no texto de edital.*

4.5 O Conjunto de dados

Os editais em formato PDF tiveram seus textos extraídos através de ferramentas de *Optical Character Recognition* (OCR). Antes disso, modelos de imagem foram utilizados para localizar as tabelas presentes nos documentos. Uma vez localizadas, as tabelas foram omitidas no processo de extração, já que a presença delas pode comprometer a capacidade dos modelos de OCR em extrair integralmente as sentenças do texto, e por

[1]Disponível em: https://docs.python.org/3/library/re.html
[2]https://huggingface.co/models
[3]https://github.com/huggingface/transformers

consequência, prejudicando o desempenho na tarefa REN, pois a identificação das entidades é altamente dependente do contexto. Trechos muito repetidos como cabeçalhos e rodapés foram omitidos dos textos extraídos.

A primeira amostra de documentos anotados foi composta por 55 editais, os textos extraídos foram submetidos à ferramenta de anotação INCEpTION [4] em arquivos de texto com uma sentença por linha. Apesar de não seguirem um padrão textual, em geral, a maior parte das ocorrências das entidades de interesse estão presentes logo no início do texto, na seção de preâmbulo, em que a licitação é apresentada em termos gerais. Considerando que um edital pode conter dezenas de páginas, incluindo seus anexos, o processo de anotação considerando a leitura integral do texto é lento e pouco produtivo em termos do número de entidades anotadas por edital.

A amostra inicial de 55 editais foi anotada por dois anotadores. Ao longo do processo de anotação, os anotadores realizaram sessões de curadoria avaliando as anotações de ambos e identificando os pontos de discrepância entre as duas anotações. As dúvidas geradas nessa fase inicial foram repassadas ao Tribunal a fim de definir os critérios de anotação. Após essa fase inicial, uma vez sanadas as dúvidas e definidos os critérios de anotação, os anotadores passaram a dividir as tarefas de anotação e curadoria, os textos anotados por um foram curados e corrigidos pelo outro anotador.

Para tornar o processo de anotação mais eficiente, foi desenvolvido um método automático de seleção de sentenças com maior probabilidade de conter alguma das entidades, comparando a similaridade semânticas entre sentenças anotadas e não-anotadas. O modelo LaBSE foi utilizado para gerar representações vetoriais das sentenças com entidades anotadas da amostra inicial. Essas sentenças foram então agrupadas por classe de entidade e a média de seus vetores foi utilizada como representação vetorial de cada uma das classes. Comparando os vetores das classes com os vetores de sentenças de editais não-anotados utilizando a similaridade do cosseno foi possível selecionar sentenças semelhantes às anotadas, e que, portanto, têm maior chance de conter entidades. A pontuação de 0.6 de similaridade do cosseno foi utilizada como limiar para seleção das sentenças. Assim, os arquivos submetidos à plataforma de anotação passaram a ser compostos por sentenças selecionadas de diversos editais.

4.5.1 Estatísticas do Conjunto de Dados

O conjunto de dados obtido é composto por 56623 sentenças e 2.077.908 *tokens*. Apenas 2,34% desses *tokens* foram anotados como entidades nomeadas, totalizando 10.416 entidades. As classes com maior número de entidades anotadas são aque-

[4] https://inception-project.github.io/

las mencionadas repetidamente no texto do edital, como *Modalidade de Licitação* e *Número/Exercício* que representam 44,62% e 25,66% do total de entidades nomeadas, respectivamente. Outras classes são mencionadas apenas uma vez em todo o texto, como *Data de Sessão* (4%), outras não são necessariamente encontradas no texto como *valor do objeto* (2,66%), já que em alguns casos essas entidades podem ser encontradas apenas em tabelas ou foram omitidas no pré-processamento realizado antes do processo de anotação.

Classe	#	%
Modalidade de Licitação	4.957	47,59
Número/Exercício	2.433	23,36
Critério de Julgamento	1.185	11,38
Objeto de Licitação	960	9,21
Data de Sessão	581	5,58
Valor do Objeto	300	2,88
Total	10.416	100

Tabela 4.1: *Distribuição das entidades anotadas por classe.*

4.6 Modelos

Considerando o domínio dos modelos derivados da arquitetura *Transformers* na tarefa REN em Língua Portuguesa [93] - especialmente os modelos da família BERT - foram selecionados para os experimentos com treinamento supervisionado a variante em Português BERTimbau (Seção 4.6.1), a variante treinada no domínio legal BERTikal (Seção 4.6.2) e o modelo que representa uma evolução da arquitetura BERT e seus métodos de treinamento, em sua versão multilíngue mDeBERTa (Seção 4.6.3).

4.6.1 BERTimbau

O modelo BERTimbau [93] é uma versão para língua portuguesa do modelo de linguagem BERT (Bidirectional Enconding Representations from Transformers) [21], originalmente desenvolvido para a língua inglesa. A arquitetura do modelo BERT deriva da arquitetura do Encoder Transformer [98]. Em relação aos modelos de linguagem anteriores, BERT apresenta a inovação de gerar representações vetoriais de palavras resultantes da combinação do contexto em ambas as direções do texto. Outra contribuição de BERT foi a definição da fase de pré-treino em que o modelo é submetido às tarefas *Mask Language Modeling* (MLM) e *Next Sentence Prediction* (NSP) em um *corpus* massivo. Na tarefa MLM o objetivo é recuperar os *tokens* que foram aleatoriamente ocultadas do texto, enquanto NSP é uma tarefa de classificação binária, ou seja, dada

uma dupla de sentenças, o objetivo é determinar se a segunda é a continuação da primeira ou não. Após o pré-treino, o modelo pode ser usado em diversas tarefas, tanto a nível de *token* quanto a nível de sentença, através da adaptação da camada de saída do modelo e o treinamento na tarefa desejada, chamado de *fine-tuning*.

Portanto, seguindo a mesma metodologia proposta por [21], o modelo BERTimbau foi pré-treinado no *corpus* em Português brWaC, composto por 3.53 milhões de documentos extraídos de páginas da Internet. O modelo obtido atingiu o estado-da-arte após o *fine-tuning* em diversas tarefas de PLN, incluindo REN no conjunto de dados HAREM.

4.6.2 BERTikal

O modelo BERTikal [74], por sua vez, é uma derivação do modelo BERTimbau, já que o modelo obtido parte do *checkpoint* obtido em [93] e passa por um pré-treino em um corpus composto por documentos jurídicos provenientes de diversas cortes do sistema judiciário brasileiro. Os experimentos realizados demonstraram uma melhora de 0.3 de F1-score em relação ao modelo BERTimbau na tarefa de classificação de texto de processos judiciais usando os *embeddings* gerados pelos modelos.

4.6.3 mDeBERTa

Após a publicação do BERT, outros trabalhos propuseram melhorias em termos das configurações de treinamento e mesmo da arquitetura *Transformers*. Dentre eles, [34] apresentou o modelo DeBERTa introduzindo duas novidades: *Disentangled Attention* (DA) e a adição da informação da posição absoluta das palavras de contexto na máscara decodificadora durante a execução da tarefa de MLM. DA utiliza um vetor para representar o conteúdo das palavras e outro para a informação da posição relativa das palavras na sentença, diferente de mecanismos de atenção anteriores em que as duas informações são incorporadas a um mesmo vetor.

Para a versão 3.0 do modelo (DebertaV3) outras inovações foram incorporadas [33]. A tarefa de pré-treino MLM foi substituída por *Replaced Token Detection* (RTD). Em RTD, dois modelos são utilizados, um como gerador e outro como discriminador. O gerador corrompe parte dos *tokens* da sequência de entrada, substituindo-os por outros *tokens*, enquanto o objetivo do discriminador é determinar, para cada um dos *tokens*, quando se trata de um *token* original ou de um *token* trocado pelo gerador.

Como extensão do trabalho da versão 3.0, também foi treinada uma versão multilinguagem o mDeBERTaV3, que superou em mais de 3% em média os modelos multilinguagem anteriores no *benchmark* para Inferência de Linguagem Natural Multilinguagem (XNLI) em 15 línguas diferentes.

4.7 Comparativo com modelos generativos

A fim de estimar a capacidade dos modelos generativos em extrair as informações dos editais de licitação. Neste estudo, utilizamos grandes Modelos de Linguagem (LLMs) para extrair e classificar informações de editais de licitação. Nossa abordagem é semelhante a uma tarefa genérica de reconhecimento de entidades nomeadas (REN), mas elimina a necessidade de treinamento específico para a tarefa e reduz a necessidade de texto anotado, uma vez que apenas alguns exemplos são usados para compor o *prompt*. O LLM é instruído a retornar uma lista de entidades extraídas do texto e classificadas por tipo. Essa lista é então usada para produzir a sequência de tags BIO (Begin, Inside e Outside), um formato comum de marcação para tarefas de classificação de *tokens*, correspondente a cada *token* da sentença (Pós-processamento na Figura 4.3). Finalmente, as previsões do LLM são comparadas com a verdade de base usando o *framework* de avaliação de rotulagem de sequência, seqeval [63]. No cenário de *Few-Shot Learning*, o processo também envolve a seleção de exemplos que compõem o *prompt* (Figura 4.3). Utilizamos precisão, *recall* e F1 score como nossas métricas de avaliação para estimar a precisão e eficácia do modelo.

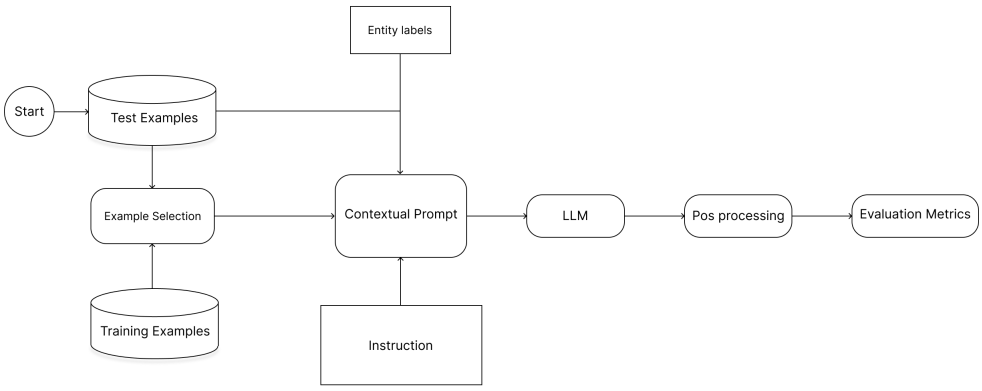


Figura 4.3: *Pipeline da abordagem de REN com Aprendizado em Contexto utilizada no cenário Few-Shot, adaptada de [67].*

4.7.1 Cenário *Zero-Shot*

No cenário *Zero-Shot*, avaliamos a capacidade dos modelos de extrair as entidades sem ajuste fino ou exemplos no *prompt*. Dois formatos, ou *templates* de *prompt* foram utilizados:

- ***template* padrão**: *este prompt contém uma descrição da tarefa de REN, a lista de tipos de entidades e instruções sobre o formato de saída*. O Apêndice A contém um exemplo de *prompt* seguindo esse padrão.
- ***template* com conceitos**: *este* template *adiciona ao* template *padrão a descrição do conceito e características de cada tipo de entidade*. O Apêndice B contém um exemplo de *prompt* seguindo esse padrão.

O objetivo de usar esses dois *templates* de *prompt* é avaliar o impacto de um *prompt* mais informativo e descritivo em comparação com um mais genérico.

4.7.2 Cenário *Few-Shot*

No cenário *Few-Shot*, o objetivo é avaliar o impacto de adicionar exemplos ao *prompt*. Cada exemplo consiste em uma sentença e sua respectiva lista de entidades. O *prompt* utilizado segue o *template com conceitos* usado no cenário *Zero-Shot* 4.3. Os exemplos foram selecionados a partir do conjunto de treinamento.

Como neste cenário, a seleção de exemplos pode ser fundamental para o sucesso desta estratégia, utilizamos uma abordagem similar à proposta em [67], com duas estratégias de seleção: **aleatória** e **similaridade semântica**. Na **seleção aleatória**, **k** exemplos de sentenças contendo entidades são selecionados aleatoriamente e incorporados ao *prompt*. Na **seleção por similaridade semântica**, os exemplos cujas sentenças são mais semanticamente similares à consulta são selecionados pela similaridade do cosseno.

CAPÍTULO 5

Experimentos e resultados

5.1 Resultados dos treinamentos supervisionados

Os modelos BERTimbau$_{\text{BASE}}$, BERTikal e mDeBERTa foram treinados de maneira supervisionada no conjunto de dados anotados. A tabela 5.1 mostra a média de f1-score obtida para cada um dos modelos nos testes de otimização de parâmetros. No espaço de busca utilizado nos testes, os parâmetros testados foram: número de épocas (10 e 20 épocas), taxa de aprendizado (1e-5, 3e-5, 4e-5, 5e-5, 6e-5, 7e-5, 8e-5 e 1e-6) e *batch size* (8, 16 e 32). Os resultados demonstram a prevalência do desempenho do modelo BERT sobre os demais, inclusive menor desvio padrão.

Modelo	**F1-score**	σ
BERTikal	0,739	0,137
mDeBERTa	0,764	0,101
BERTimbau	**0,774**	**0,056**

Tabela 5.1: *Média do desempenho dos modelos nos testes de otimização de parâmetros.*

Embora observando a métrica geral o modelo BERT seja superior, a análise sobre as métricas por classe de entidades para os melhores resultados de cada modelo (Figura 5.1), revela que para algumas classes os outros modelos foram superiores. Com destaque para mDeBERTa, melhor nas classes Data de Sessão, Número/Exercício e Valor do Objeto.

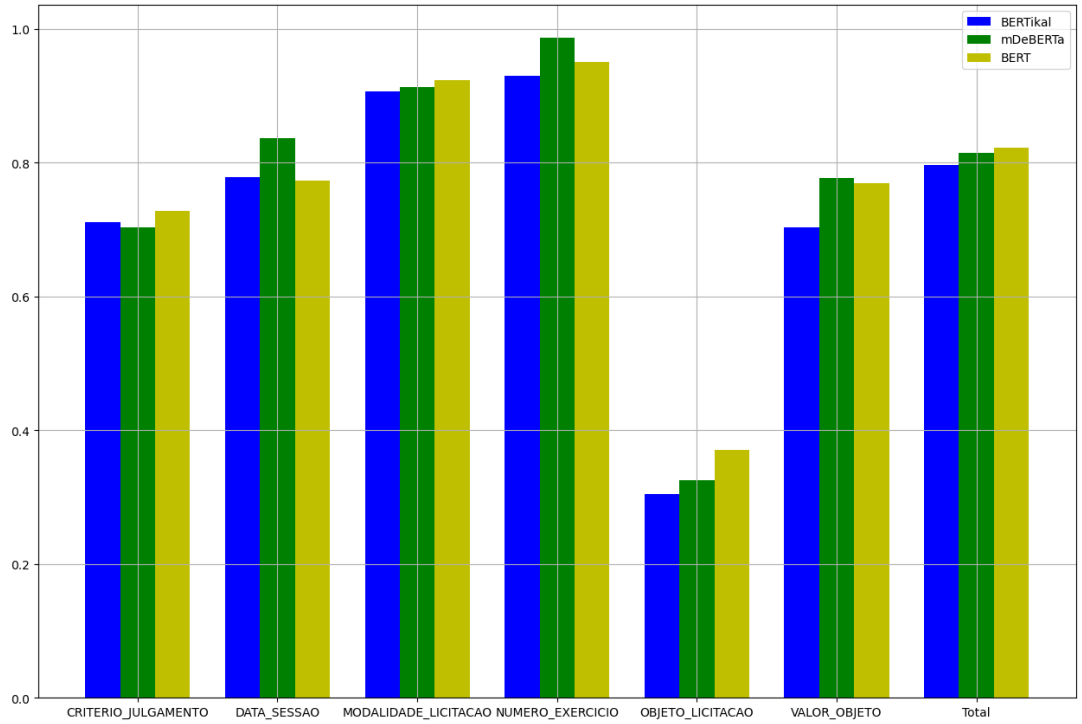


Figura 5.1: *Métricas por classe de entidade para os melhores resultados de cada modelo.*

A classe mais difícil para extração é o Objeto da Licitação. Por ser uma informação composta por múltiplas palavras, a definição do início, e principalmente do fim da entidade é bastante subjetiva para essa classe, inclusive no processo de anotação, portanto entidades cujos *tokens* são identificadas em parte deterioram as métricas. Já as classes mais padronizadas como Modalidade e Número/Exercício têm F1-score acima de 0.8 para todos os modelos. Outras classes sofrem com falsos negativos, é o caso de Critério de Julgamento, em que as palavras que compõem a entidade podem estar presentes em outros contextos no documento. Já para Valor de Objeto e Data de Sessão, a presença no texto de outras datas e valores monetários dificultam a identificação da entidade correta.

5.2 *Zero-shot* e *Few-shot Learning* com LLMs

Nesta seção, apresentamos uma descrição completa dos dois cenários de experimentos: zero-shot e few-shot. Para cada cenário, revelamos os resultados e oferecemos uma análise comparativa entre os modelos testados.

5.2.1 Cenário Zero-Shot

No cenário Zero-Shot, avaliamos a capacidade dos modelos de extrair as entidades sem *fine-tuning* ou exemplos no *prompt*. Dois tipos de *templates* de *prompt* foram

utilizados nos experimentos:

- **Template Padrão**: este *prompt* contém uma descrição da tarefa de NER, a lista de tipos de entidades e instruções sobre o formato de saída.
- **Template com Conceitos**: este template adiciona ao Template Padrão a descrição do conceito e características de cada tipo de entidade.

O objetivo de utilizar esses dois *templates* é avaliar o impacto de um *prompt* mais informativo e descritivo em comparação a um mais genérico.

De modo geral, o modelo gpt-4o-mini foi o melhor em ambos os *templates*, com F1-score de 0,4069 (Template Padrão) e 0,39 (Template com Conceitos), seguido pelo llama3-70b-8192 com 0,4789 e 0,3062, respectivamente. Com exceção do tipo de entidade MODALIDADE_LICITACAO, o gpt-4o-mini superou o llama3-70b-8192 em todas as demais entidades.

Modelos	**Template Padrão**			**Template com Conceitos**		
	Precisão	Recall	F1-Score	Precisão	Recall	F1-Score
gpt-3.5-turbo	0,1313	0,3058	0,1837	0,2004	0,2811	0,2340
gpt-4o-mini	0,4422	**0,3768**	**0,4069**	**0,4789**	0,3290	0,3900
llama3-8b-8192	0,0592	0,1478	0,0846	0,1158	0,1783	0,1404
llama3-70b-8192	0,1405	0,2928	0,1898	0,2651	0,3623	0,3062

Tabela 5.2: *Resultados da avaliação no cenário zero-shot para ambos os templates: padrão e com conceitos.*

Os resultados na Tabela 5.2 revelam uma melhora mais significativa para os modelos LLaMA ao usar o *template* com conceitos, enquanto nos modelos GPT houve ganho em precisão, mas com certa perda na Recall. Isso é evidente no desempenho do gpt-4o-mini para a entidade NUMERO_EXERCICIO, por exemplo. Embora tenha ocorrido um ganho em precisão, o *prompt* mais restritivo gerou um maior número de falsos negativos, resultando em um desempenho geral ligeiramente inferior ao do *template* padrão.

A Figura 5.2 também mostra que os tipos de entidade mais desafiadores são VALOR_OBJETO e OBJETO_LICITACAO, ambos relacionados ao conceito de objeto do processo licitatório. Embora o *Template com Conceitos* contenha a explicação desse conceito, os resultados não apresentaram grande melhora para nenhum dos modelos testados.

5.2.2 Cenário Few-Shot

No cenário Few-Shot, avaliamos o impacto de adicionar exemplos ao prompt. Cada exemplo consiste em uma sentença e sua lista correspondente de entidades. O *prompt* utilizado segue o *Template com Conceitos* do cenário Zero-Shot.

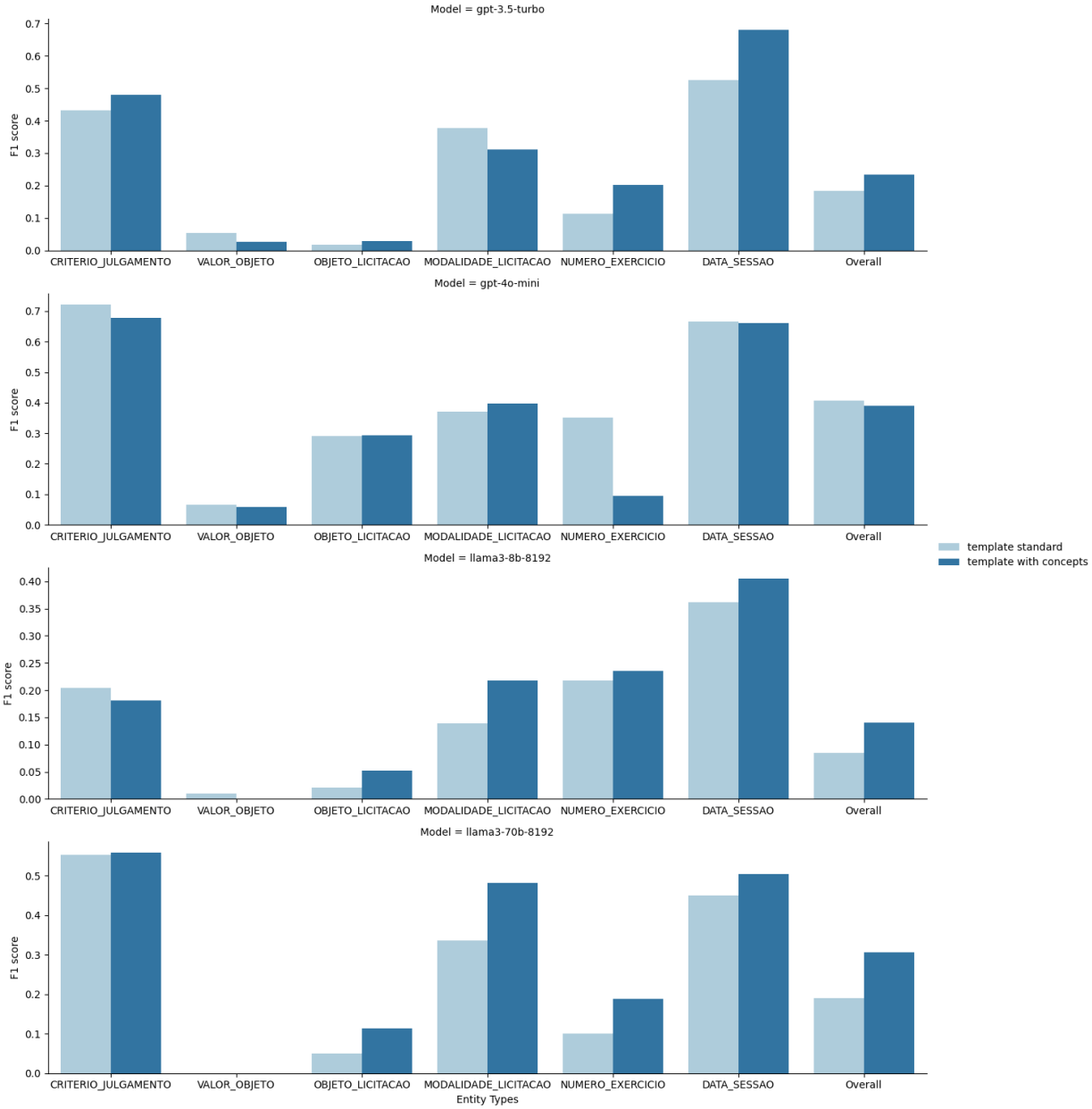


Figura 5.2: *Métricas de avaliação por tipo de entidade no cenário Zero-Shot.*

Os exemplos foram selecionados a partir do conjunto de treinamento de 27.971 sentenças anotadas manualmente. Similarmente ao proposto em [67], utilizamos diferentes estratégias de seleção nos experimentos: **aleatória** e **similaridade semântica**. Na seleção aleatória, **k** exemplos de sentenças com entidades são escolhidos aleatoriamente e incorporados ao prompt. Na seleção por similaridade semântica, os exemplos mais semanticamente próximos da consulta são selecionados por similaridade de cosseno. As embeddings utilizadas foram geradas pelo modelo LaBSE [26]. Testamos entre 1 e 10 exemplos.

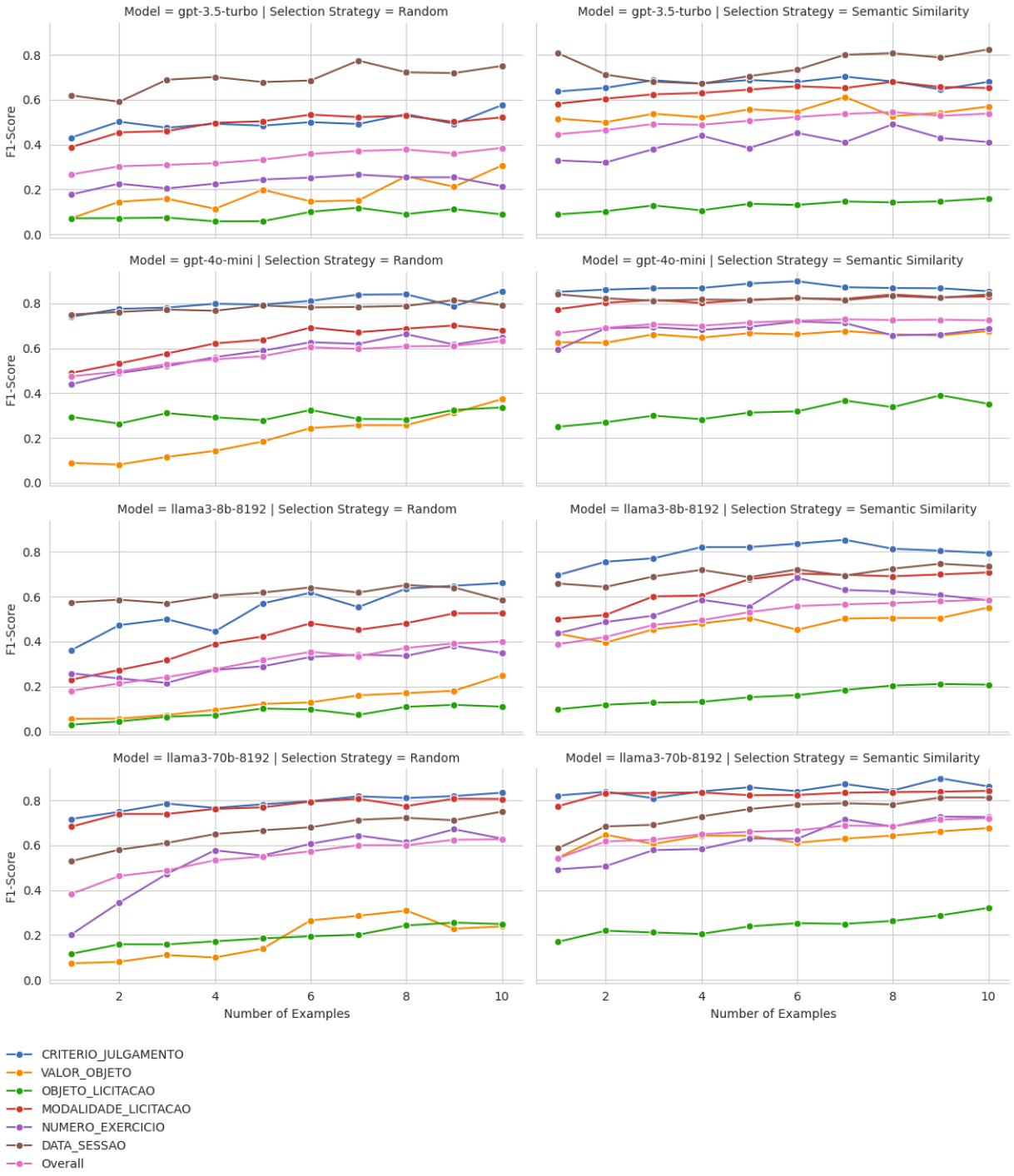


Figura 5.3: *Desempenho dos modelos no cenário Few-shot conforme o número de exemplos.*

Análise Comparativa do Cenário Few-Shot

O gpt-4o-mini alcançou um F1-score de 0,7234 no desempenho geral com 10 exemplos selecionados por similaridade semântica, seguido de perto pelo llama3-70b-8192 com F1-score de 0,720 na mesma configuração. O aumento do número de exemplos no *prompt* parece beneficiar mais os modelos LLaMA do que os modelos GPT. Ao usar a

similaridade semântica como estratégia de seleção, por exemplo, o desempenho geral do gpt-4o-mini se manteve relativamente estável, ligeiramente acima de 0,70 de F1-score entre 2 e 10 exemplos, enquanto o llama3-70b-8192 começou com 0,5421 com um exemplo e superou a marca de 0,70 apenas com 9 exemplos.

Seleção Aleatória versus Seleção por Similaridade Semântica

A diferença de desempenho entre as duas estratégias de seleção é mais pronunciada para a entidade VALOR_OBJETO (Figura 5.3). Como a entidade VALOR_OBJETO aparece isoladamente na maioria das sentenças, enquanto outros tipos têm maior chance de aparecer juntos na mesma sentença, a seleção por similaridade semântica foi crucial para encontrar exemplos mais significativos.

5.2.3 Comparativo entre as abordagens

Comparando os melhores resultados obtidos pelo treinamento supervisionado e pelas LLMs, o desempenho de ambos foi semelhante. A Figura 5.4 apresenta o F1-score obtido por tipo de entidade para o gpt4-mini no cenário com 10 exemplos selecionados por similaridade semântica e o modelo BERTimbau. Apesar dos desempenhos serem similares, o BERTimbau tem uma pequena vantagem para os tipos mais padronizados como Modalidade de Licitação e Número/Exercício.

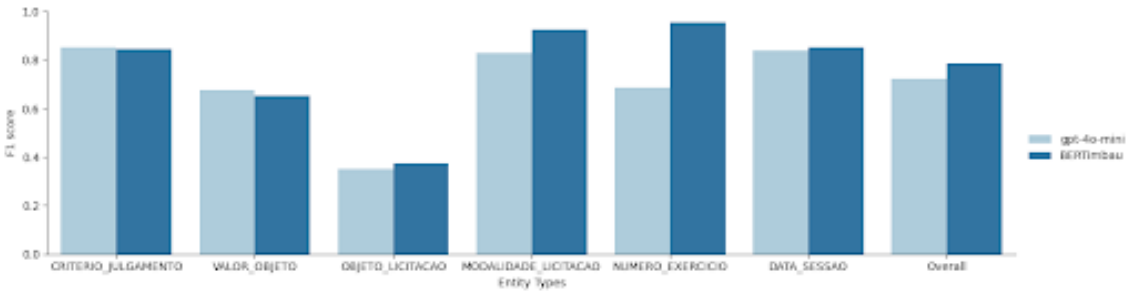


Figura 5.4: *Comparativo entre abordagens BERTimbau e GPT 4 omni mini com 10 exemplos.*

CAPÍTULO 6

Conclusão

Nesse trabalho foi apresentado um estudo a aplicação de modelos de linguagem natural (LLMs) para extração de informações em editais de licitação, utilizando a tarefa de Reconhecimento de Entidades Nomeadas (REN). O objetivo principal foi explorar técnicas que otimizassem o processo de anotação e extração de entidades específicas, minimizando o esforço manual e maximizando a precisão dos modelos em um conjunto de dados de editais.

O trabalho foi guiado pelas seguintes perguntas de pesquisa:

Qual o impacto de diferentes *templates* de prompt e estratégias de exemplo no desempenho dos modelos em cenários *Zero-Shot* e *Few-Shot Learning* ? Quais classes de entidade apresentam maior dificuldade para extração automática e por quê?

Para responder a essas questões, os experimentos foram divididos em três partes principais: (i) construção e anotação do conjunto de dados, (ii) avaliação do desempenho dos modelos de REN em configurações supervisionadas, e (iii) experimentos de Zero-Shot e Few-Shot, explorando diferentes configurações de *prompts* e estratégias de seleção de exemplos.

No primeiro estágio, foi desenvolvido um processo de anotação que utilizou um método semiautomático para selecionar sentenças mais prováveis de conterem entidades. A seleção foi baseada em similaridade semântica entre as sentenças anotadas e as não-anotadas. Essa técnica melhorou a eficiência da anotação ao reduzir o número de sentenças irrelevantes e ajudou a criar um conjunto de dados com distribuição balanceada entre as classes de entidade.

No segundo estágio, os modelos BERTimbau, BERTikal e mDeBERTa foram treinados no conjunto de dados anotado. Os resultados indicaram que o modelo BERTimbau obteve o melhor desempenho geral, especialmente nas classes mais padronizadas, como Modalidade de Licitação e Número/Exercício, atingindo F1-scores acima de 0,8. Entretanto, classes mais complexas e com menor regularidade textual, como Objeto de Licitação e Valor do Objeto, apresentaram maior dificuldade de extração, devido à variabilidade de formato e localização dessas informações nos editais.

No terceiro e último estágio, os modelos foram testados em cenários *Zero-Shot* e *Few-Shot*. Os experimentos indicaram que, embora *prompts* com descrição de conceitos tenha elevado o desempenho, a adição de exemplos no formato Few-Shot foi fundamental para aumentar a precisão, principalmente nas classes mais complexas. A seleção dos exemplos com base na similaridade semântica teve um efeito positivo no desempenho dos modelos, especialmente para classes de entidade como Valor do Objeto, em que a presença de exemplos específicos no *prompt* ajudou os modelos a identificar corretamente entidades semelhantes nas novas sentenças.

As análises indicam que, para modelos de REN aplicados a editais de licitação, a combinação de *prompts* ricos em exemplos semanticamente similares foi eficaz em melhorar a precisão do processo de extração de entidades. No contexto de tarefas REN mais complexas e com baixo padrão textual, o uso de modelos de linguagem de última geração, somado a técnicas de seleção semântica, revela-se promissor, permitindo avanços na automação de sistemas de extração de informações estruturadas em documentos administrativos.

Respondendo às questões de pesquisa (Capítulo 1), podemos afirmar que:

Questão 1: Os modelos treinados no *corpus* anotado foram capazes de extrair os dados, com F1-score acima de 0.80 para maioria dos tipos de entidade. Embora a métrica para o tipo Objeto de Licitação seja inferior a 0.4, na prática os verdadeiros positivos identificados pelos modelos são suficientes para a identificação dos objetos de licitação na aplicação do sistema proposto no Tribunal.

Questão 2: Sim, as LLMs ao menos quando auxiliadas pelas técnicas de ICL podem alcançar desempenho semelhante aos modelos treinados via treinamento supervisionado.

Questão 3: Quanto aos *templates* de *prompts* usados nos experimentos, o impacto da incorporação dos conceitos dos tipos de entidade foi pequeno, já que permitiu maior precisão, mas a abrangência foi reduzida. Dentre os elementos que compõem o *prompt*, aquele com maior impacto positivo para extração das informações foi o uso de exemplos, em especial exemplos cujas sentenças sejam similares à sentença de entrada.

Questão 4: A classe mais desafiadora é o Objeto de Licitação, por ser uma sentença por vez extensa e com grande variedade de conteúdo. Apesar disso, analisando a extração por edital, os modelos foram capazes de extrair ao menos uma ocorrência correta de Objeto de Licitação.

Por fim, os resultados deste estudo contribuem para o desenvolvimento de métodos mais eficientes de extração de informações em documentos públicos, com potenciais aplicações em áreas como transparência governamental e monitoramento de licitações, ajudando a viabilizar sistemas mais precisos e acessíveis para análise de dados em grande escala.

6.1 Limitações e Trabalhos Futuros

O trabalho desenvolvido para os cenários de REN em *Zero-Shot* e *Few-shot* Learning, se limitou a explorar os LLMs generativos multilinguagem que incluem o Português. Para um melhor entendimento da eficácia dos métodos utilizados, um comparativo mais completo deveria incluir outros métodos recentes de Zero-Shot para REN, como GliNER [105], baseado na arquitetura BERT, como também incluir LLMs especializadas em Português, como Sabiá [72].

Para trabalhos futuros, há a possibilidade de explorar a adaptação desses modelos ao domínio legal através de treinamento supervisionado na tarefa de REN, ou mesmo outras tarefas de extração de informação. Assim, seria necessário um conjunto de treinamento suficientemente diverso em termos do números de tipos de entidades nomeadas, para produzir modelos capazes de generalizar tarefa para diferentes tipos de entidade e domínios. Uma alternativa seria utilizar como dados de treinamento a compilação do *corpora* disponível para REN em Português, semelhante ao conjunto de dados UniNER [109] que foi desenvolvido para a Língua Inglesa. Outro caminho viável é a criação de *corpus* anotados automaticamente por LLMs para enriquecer o conjunto de treinamento. Por fim, o desenvolvimento de um *benchmark* mais abrangente, que inclua outros subdomínios de textos do Direito seria fundamental para a correta comparação de novas técnicas de REN.

Referências

[1] AGUILAR, G.; MAHARJAN, S.; LÓPEZ-MONROY, A. P.; SOLORIO, T. **A multi-task approach for named entity recognition in social media data**. *arXiv preprint arXiv:1906.04135*, 2019.

[2] AKBIK, A.; BERGMANN, T.; BLYTHE, D.; RASUL, K.; SCHWETER, S.; VOLLGRAF, R. **Flair: An easy-to-use framework for state-of-the-art nlp**. In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, p. 54–59, 2019.

[3] ALBUQUERQUE, H. O.; COSTA, R.; SILVESTRE, G.; SOUZA, E.; DA SILVA, N. F.; VITÓRIO, D.; MORIYAMA, G.; MARTINS, L.; SOEZIMA, L.; NUNES, A.; OTHERS. **Ulyssesner-br: a corpus of brazilian legislative documents for named entity recognition**. In: *International Conference on Computational Processing of the Portuguese Language*, p. 3–14. Springer, 2022.

[4] ALBUQUERQUE, H. O.; SOUZA, E.; GOMES, C.; PINTO, M. H. D. C.; RICARDO FILHO, P.; COSTA, R.; LOPES, V. T. D. M.; DA SILVA, N. F.; DE CARVALHO, A. C.; OLIVEIRA, A. L. **Named entity recognition: a survey for the portuguese language**. *Procesamiento del Lenguaje Natural*, 70:171–185, 2023.

[5] ALLES, V. J. **Construção de um corpus para extrair entidades nomeadas do diário oficial da união utilizando aprendizado supervisionado**. 2018.

[6] ALMEIDA, F.; XEXÉO, G. **Word embeddings: A survey**. *arXiv preprint arXiv:1901.09069*, 2019.

[7] AMARAL, D.; VIEIRA, R. **Nerp-crf: a tool for the named entity recognition using conditional random fields**, 2014.

[8] AMARAL, D.; COLLOVINI, S.; FIGUEIRA, A.; VIEIRA, R.; GONZALEZ, M. **Processo de construção de um corpus anotado com entidades geológicas visando ren (building an annotated corpus with geological entities for ner)[in portuguese]**. In: *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, p. 63–72, 2017.

[9] BATISTA, H. H.; NASCIMENTO, A. C.; MELO, R. F.; MIRANDA, P. B.; MALDONADO, I. W.; COELHO FILHO, J. L. **A comparative analysis of text embedding approach to extract named entities in portuguese legal documents**. In: *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, p. 221–232. SBC, 2021.

[10] BONIFACIO, L. H.; VILELA, P. A.; LOBATO, G. R.; FERNANDES, E. R. **A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese**. In: *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, p. 648–662. Springer, 2020.

[11] CASTRO, P. V. Q. D.; OTHERS. **Aprendizagem profunda para reconhecimento de entidades nomeadas em domínio jurídico**. 2019.

[12] CHANG, Y.; WANG, X.; WANG, J.; WU, Y.; YANG, L.; ZHU, K.; CHEN, H.; YI, X.; WANG, C.; WANG, Y.; OTHERS. **A survey on evaluation of large language models**. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[13] CHOROWSKI, J.; BAHDANAU, D.; SERDYUK, D.; CHO, K.; BENGIO, Y. **Attention-based models for speech recognition**. *arXiv preprint arXiv:1506.07503*, 2015.

[14] CHOWDHARY, K.; CHOWDHARY, K. **Natural language processing**. *Fundamentals of artificial intelligence*, p. 603–649, 2020.

[15] COATES-STEPHENS, S. **The analysis and acquisition of proper names for the understanding of free text**. *Computers and the Humanities*, 26:441–456, 1992.

[16] COLLOBERT, R.; WESTON, J.; BOTTOU, L.; KARLEN, M.; KAVUKCUOGLU, K.; KUKSA, P. **Natural language processing (almost) from scratch**. *J. Mach. Learn. Res.*, 12:2493–2537, Nov. 2011.

[17] CONSOLI, B.; SANTOS, J.; GOMES, D.; CORDEIRO, F.; VIEIRA, R.; MOREIRA, V. **Embeddings for named entity recognition in geoscience portuguese literature**. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 4625–4630, 2020.

[18] COSTA, R.; ALBUQUERQUE, H. O.; SILVESTRE, G.; SILVA, N. F. F.; SOUZA, E.; VITÓRIO, D.; NUNES, A.; SIQUEIRA, F.; PEDRO TARREGA, J.; VITOR BEINOTTI, J.; OTHERS. **Expanding ulyssesner-br named entity recognition corpus with informal user-generated text**. In: *Progress in Artificial Intelligence: 21st EPIA*

Conference on Artificial Intelligence, EPIA 2022, Lisbon, Portugal, August 31–September 2, 2022, Proceedings, p. 767–779. Springer, 2022.

[19] DE AQUINO SILVA, R.; DA SILVA, L.; DUTRA, M. L.; DE ARAUJO, G. M. **A new entity extraction model based on journalistic brazilian portuguese language to enhance named entity recognition**. In: *International Conference on Data and Information in Online*, p. 53–63. Springer, 2020.

[20] DE MELLO, C. A. B. **O edital nas licitações**. *Revista de Direito Administrativo*, 131:281–299, 1978.

[21] DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. **Bert: Pre-training of deep bidirectional transformers for language understanding**. *arXiv preprint arXiv:1810.04805*, 2018.

[22] DO AMARAL, D. O.; FONSECA, E.; LOPES, L.; VIEIRA, R. **Comparing nerp-crf with publicly available portuguese named entities recognition tools**. In: *Computational Processing of the Portuguese Language: 11th International Conference, PROPOR 2014, São Carlos/SP, Brazil, October 6-8, 2014. Proceedings 11*, p. 244–249. Springer, 2014.

[23] DONG, Q.; LI, L.; DAI, D.; ZHENG, C.; MA, J.; LI, R.; XIA, H.; XU, J.; WU, Z.; LIU, T.; OTHERS. **A survey on in-context learning**. *arXiv preprint arXiv:2301.00234*, 2022.

[24] EDDY, S. R. **Hidden markov models**. *Current opinion in structural biology*, 6(3):361–365, 1996.

[25] ETZIONI, O.; CAFARELLA, M.; DOWNEY, D.; POPESCU, A.-M.; SHAKED, T.; SODERLAND, S.; WELD, D. S.; YATES, A. **Unsupervised named-entity extraction from the web: An experimental study**. *Artificial intelligence*, 165(1):91–134, 2005.

[26] FENG, F.; YANG, Y.; CER, D.; ARIVAZHAGAN, N.; WANG, W. **Language-agnostic bert sentence embedding**. *arXiv preprint arXiv:2007.01852*, 2020.

[27] FONSECA, E. B.; ANTONITSCH, A.; COLLOVINI, S.; AMARAL, D.; VIEIRA, R.; FIGUEIRA, A. **Summ-it++: an enriched version of the summ-it corpus**. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 2047–2051, 2016.

[28] FREITAS, C.; CARVALHO, P.; GONÇALO OLIVEIRA, H.; MOTA, C.; SANTOS, D. **Second harem: advancing the state of the art of named entity recognition in**

portuguese. in quot. *Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan Odijk; Stelios Piperidis*, 2010.

[29] GONÇALVES, M.; COHEUR, L.; BAPTISTA, J.; MINEIRO, A. **Avaliação de recursos computacionais para o português**. *Linguamática*, 12(2):51–68, 2021.

[30] GOODFELLOW, I. **Deep learning**, 2016.

[31] GRISHMAN, R.; SUNDHEIM, B. M. **Message understanding conference-6: A brief history**. In: *COLING 1996 volume 1: The 16th international conference on computational linguistics*, 1996.

[32] GUPTA, S.; GARDNER, M.; SINGH, S. **Coverage-based example selection for in-context learning**. *arXiv preprint arXiv:2305.14907*, 2023.

[33] HE, P.; GAO, J.; CHEN, W. **Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing**. *arXiv preprint arXiv:2111.09543*, 2021.

[34] HE, P.; LIU, X.; GAO, J.; CHEN, W. **Deberta: Decoding-enhanced bert with disentangled attention**. *arXiv preprint arXiv:2006.03654*, 2020.

[35] HEARST, M. A.; DUMAIS, S. T.; OSUNA, E.; PLATT, J.; SCHOLKOPF, B. **Support vector machines**. *IEEE Intelligent Systems and their applications*, 13(4):18–28, 1998.

[36] HIGUCHI, S.; FREITAS, C.; CUCONATO, B.; RADEMAKER, A. **Text mining for history: first steps on building a large dataset**. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[37] HUANG, Z.; XU, W.; YU, K. **Bidirectional lstm-crf models for sequence tagging**. *arXiv preprint arXiv:1508.01991*, 2015.

[38] JOULIN, A.; GRAVE, E.; BOJANOWSKI, P.; MIKOLOV, T. **Bag of tricks for efficient text classification**. *arXiv preprint arXiv:1607.01759*, 2016.

[39] JÚNIOR, J. C. **Natureza das decisões do tribunal de contas**. *Revista de Direito Administrativo*, 166:1–16, 1986.

[40] KALCHBRENNER, N.; BLUNSOM, P. **Recurrent continuous translation models**. In: *Proceedings of the 2013 conference on empirical methods in natural language processing*, p. 1700–1709, 2013.

[41] KRIPKE, S. **Naming and necessity**, 1980.

[42] KURU, O.; CAN, O. A.; YURET, D. **Charner: Character-level named entity recognition**. In: *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, p. 911–921, 2016.

[43] LAFFERTY, J.; MCCALLUM, A.; PEREIRA, F.; OTHERS. **Conditional random fields: Probabilistic models for segmenting and labeling sequence data**. In: *Icml*, volume 1, p. 3. Williamstown, MA, 2001.

[44] LAMPLE, G. **Neural architectures for named entity recognition**. *arXiv preprint arXiv:1603.01360*, 2016.

[45] LI, J.; SUN, A.; HAN, J.; LI, C. **A survey on deep learning for named entity recognition**. *IEEE transactions on knowledge and data engineering*, 34(1):50–70, 2020.

[46] LI, J.; SUN, A.; JOTY, S. R. **Segbot: A generic neural text segmentation model with pointer network.** In: *IJCAI*, p. 4166–4172, 2018.

[47] LI, P.-H.; DONG, R.-P.; WANG, Y.-S.; CHOU, J.-C.; MA, W.-Y. **Leveraging linguistic structures for named entity recognition with bidirectional recursive neural networks**. In: *Proceedings of the 2017 conference on empirical methods in natural language processing*, p. 2664–2669, 2017.

[48] LIU, T.; YAO, J.-G.; LIN, C.-Y. **Towards improving neural named entity recognition with gazetteers**. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*, p. 5301–5307, 2019.

[49] LOPES, F.; TEIXEIRA, C.; GONÇALO OLIVEIRA, H. **Named entity recognition in portuguese neurology text using crf**. In: *Progress in Artificial Intelligence: 19th EPIA Conference on Artificial Intelligence, EPIA 2019, Vila Real, Portugal, September 3–6, 2019, Proceedings, Part I 19*, p. 336–348. Springer, 2019.

[50] LUONG, M.-T.; PHAM, H.; MANNING, C. D. **Effective approaches to attention-based neural machine translation**. *arXiv preprint arXiv:1508.04025*, 2015.

[51] LUZ DE ARAUJO, P. H.; DE CAMPOS, T. E.; DE OLIVEIRA, R. R.; STAUFFER, M.; COUTO, S.; BERMEJO, P. **Lener-br: a dataset for named entity recognition in brazilian legal text**. In: *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, p. 313–323. Springer, 2018.

[52] MA, X. **End-to-end sequence labeling via bi-directional lstm-cnns-crf**. *arXiv preprint arXiv:1603.01354*, 2016.

[53] MATOS, E.; RODRIGUES, M.; TEIXEIRA, A. **Towards the automatic creation of NER systems for new domains**. In: Gamallo, P.; Claro, D.; Teixeira, A.; Real, L.; Garcia, M.; Oliveira, H. G.; Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, p. 218–227, Santiago de Compostela, Galicia/Spain, Mar. 2024. Association for Computational Lingustics.

[54] MENEZES, D.; MILIDIU, R.; SAVARESE, P. **Building a massive corpus for named entity recognition using free open data sources**. In: *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, p. 6–11. IEEE, 2019.

[55] MIKHEEV, A.; MOENS, M.; GROVER, C. **Named entity recognition without gazetteers**. In: *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, p. 1–8, 1999.

[56] MIKOLOV, T. **Efficient estimation of word representations in vector space**. *arXiv preprint arXiv:1301.3781*, 2013.

[57] MIKOLOV, T.; SUTSKEVER, I.; CHEN, K.; CORRADO, G. S.; DEAN, J. **Distributed representations of words and phrases and their compositionality**. *Advances in neural information processing systems*, 26, 2013.

[58] MIRANDA, N.; RAMINHOS, R.; SEABRA, P.; SEQUEIRA, J.; GONÇALVES, T.; QUARESMA, P. **Named entity recognition using machine learning techniques**. In: *Epia-11, 15th portuguese conference on artificial intelligence*, p. 818–831, 2011.

[59] MOHIT, B. **Named entity recognition**. In: *Natural language processing of semitic languages*, p. 221–245. Springer, 2014.

[60] MOREIRA, F.; VIEIRA, R. **Aplicação de reconhecimento de entidades nomeadas em investigação de crimes financeiros**. In: *Proceedings of the 2nd Symposium in information and human language technology*, p. 134–143, 2019.

[61] MOTA, C. C.; NASCIMENTO, A. C.; MIRANDA, P. B.; MELLO, R. F.; MALDONADO, I. W.; COELHO FILHO, J. L. **Reconhecimento de entidades nomeadas em documentos jurídicos em português utilizando redes neurais**. In: *Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*, p. 130–140. SBC, 2021.

[62] NADEAU, D.; SEKINE, S. **A survey of named entity recognition and classification**. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[63] NAKAYAMA, H. **seqeval: A python framework for sequence labeling evaluation**, 2018. Software available from https://github.com/chakki-works/seqeval.

[64] NASAR, Z.; JAFFRY, S. W.; MALIK, M. K. **Named entity recognition and relation extraction: State-of-the-art**. *ACM Computing Surveys (CSUR)*, 54(1):1–39, 2021.

[65] NETO, J. R. C.; FALEIROS, T. D. P. **Deep active-self learning applied to named entity recognition**. In: *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, p. 405–418. Springer, 2021.

[66] NGUYEN, T. H.; SIL, A.; DINU, G.; FLORIAN, R. **Toward mention detection robustness with recurrent neural networks**. *arXiv preprint arXiv:1602.07749*, 2016.

[67] NUNES, R. O.; SPRITZER, A.; FREITAS, C. D. S.; BALREIRA, D. **Out of sesame street: A study of portuguese legal named entity recognition through in-context learning**. In: *Proceedings of the 26th International Conference on Enterprise Information Systems - Volume 1: ICEIS*, p. 477–489. INSTICC, SciTePress, 2024.

[68] PENNINGTON, J.; SOCHER, R.; MANNING, C. D. **Glove: Global vectors for word representation**. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, p. 1532–1543, 2014.

[69] PETASIS, G.; CUCCHIARELLI, A.; VELARDI, P.; PALIOURAS, G.; KARKALETSIS, V.; SPYROPOULOS, C. D. **Automatic adaptation of proper noun dictionaries through cooperation of machine learning and probabilistic methods**. In: *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, p. 128–135, 2000.

[70] PETERS, M. E.; NEUMANN, M.; IYYER, M.; GARDNER, M.; CLARK, C.; LEE, K.; ZETTLEMOYER, L. **Deep contextualized word representations. corr abs/1802.05365 (2018)**. *arXiv preprint arXiv:1802.05365*, 1802.

[71] PETERS, M. E.; NEUMANN, M.; ZETTLEMOYER, L.; YIH, W.-T. **Dissecting contextual word embeddings: Architecture and representation**. *arXiv preprint arXiv:1808.08949*, 2018.

[72] PIRES, R.; ABONIZIO, H.; ALMEIDA, T. S.; NOGUEIRA, R. **Sabiá: Portuguese large language models**. In: Naldi, M. C.; Bianchi, R. A. C., editors, *Intelligent Systems*, p. 226–240, Cham, 2023. Springer Nature Switzerland.

[73] PIROVANI, J.; OLIVEIRA, E. **Portuguese named entity recognition using conditional random fields and local grammars**. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[74] POLO, F. M.; MENDONÇA, G. C. F.; PARREIRA, K. C. J.; GIANVECHIO, L.; CORDEIRO, P.; FERREIRA, J. B.; DE LIMA, L. M. P.; DO AMARAL MAIA, A. C.; VICENTE, R. **Legalnlp-natural language processing methods for the brazilian legal language**. In: *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional*, p. 763–774. SBC, 2021.

[75] QIU, X.; SUN, T.; XU, Y.; SHAO, Y.; DAI, N.; HUANG, X. **Pre-trained models for natural language processing: A survey**. *CoRR*, abs/2003.08271, 2020.

[76] QUINLAN, J. R. **Induction of decision trees**. *Machine learning*, 1:81–106, 1986.

[77] QUINTA DE CASTRO, P. V.; FÉLIX FELIPE DA SILVA, N.; DA SILVA SOARES, A. **Portuguese named entity recognition using lstm-crf**. In: *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, p. 83–92. Springer, 2018.

[78] RADFORD, A. **Improving language understanding by generative pre-training**. 2018.

[79] RAU, L. F. **Extracting company names from text**. In: *Proceedings the seventh IEEE conference on artificial intelligence application*, p. 29–30. IEEE Computer Society, 1991.

[80] RAVIN, Y.; WACHOLDER, N. **Extracting names from natural-language text**. Citeseer, 1997.

[81] REYES, D. D. L.; TRAJANO, D.; MANSSOUR, I. H.; VIEIRA, R.; BORDINI, R. H. **Entity relation extraction from news articles in portuguese for competitive intelligence based on bert**. In: *Brazilian Conference on Intelligent Systems*, p. 449–464. Springer, 2021.

[82] RODRÍGUEZ, M. M.; BEZERRA, B. L. D. **Processamento de linguagem natural para reconhecimento de entidades nomeadas em textos jurídicos de atos administrativos (portarias)**. *Revista de Engenharia e Pesquisa Aplicada*, 5(1):67–77, 2020.

[83] SAMPAIO, V.A., M. F. P. S. G. C.; HISSA, L. **A brief survey of deep learning based methods against opennlp namefinder for named entity recognition on**

portuguese literary texts. *In Proceedings of XII Symposium in Information and Human Language Technology and Collocates Events (STIL 2019)*, 2019.

[84] SANTOS, C. N. D.; GUIMARAES, V. **Boosting named entity recognition with neural character embeddings**. *arXiv preprint arXiv:1505.05008*, 2015.

[85] SANTOS, D.; CARDOSO, N. **A golden resource for named entity recognition in portuguese**. In: *International workshop on computational processing of the portuguese language*, p. 69–79. Springer, 2006.

[86] SANTOS, J.; CONSOLI, B.; DOS SANTOS, C.; TERRA, J.; COLLONINI, S.; VIEIRA, R. **Assessing the impact of contextual embeddings for portuguese named entity recognition**. In: *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, p. 437–442. IEEE, 2019.

[87] SEKINE, S.; NOBATA, C. **Definition, dictionaries and tagger for extended named entity hierarchy.** In: *LREC*, p. 1977–1980. Lisbon, Portugal, 2004.

[88] SETTLES, B. **Biomedical named entity recognition using conditional random fields and rich feature sets**. In: *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*, p. 107–110, 2004.

[89] SHARNAGAT, R. **Named entity recognition: A literature survey**. *Center For Indian Language Technology*, p. 1–27, 2014.

[90] SHEN, Y.; YUN, H.; LIPTON, Z. C.; KRONROD, Y.; ANANDKUMAR, A. **Deep active learning for named entity recognition**. *arXiv preprint arXiv:1707.05928*, 2017.

[91] SILVA, D. F.; SILVA, A. M. E.; LOPES, B. M.; JOHANSSON, K. M.; ASSI, F. M.; DE JESUS, J. T.; MAZO, R. N.; LUCRÉDIO, D.; CASELI, H. M.; REAL, L. **Named entity recognition for brazilian portuguese product titles**. In: *Intelligent Systems: 10th Brazilian Conference, BRACIS 2021, Virtual Event, November 29–December 3, 2021, Proceedings, Part II 10*, p. 526–541. Springer, 2021.

[92] SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. **Portuguese named entity recognition using bert-crf**. *arXiv preprint arXiv:1909.10649*, 2019.

[93] SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. **Bertimbau: pretrained bert models for brazilian portuguese**. In: *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*, p. 403–417. Springer, 2020.

[94] SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. **Bert models for brazilian portuguese: Pretraining, evaluation and tokenization analysis**. *Applied Soft Computing*, 149:110901, 2023.

[95] STRUBELL, E.; VERGA, P.; BELANGER, D.; MCCALLUM, A. **Fast and accurate entity recognition with iterated dilated convolutions. arxiv 2017**. *arXiv preprint arXiv:1702.02098*, 2017.

[96] SUTSKEVER, I.; VINYALS, O.; LE, Q. V. **Sequence to sequence learning with neural networks**. In: *Advances in neural information processing systems*, p. 3104–3112, 2014.

[97] TETÉO, L.; MOURA, P.; SOARES, E. F. D. S.; CAMPOS, C. A. V. **Um framework de extraçao e etiquetamento de informaçoes de trânsito**. In: *Anais do XVIII Workshop em Desempenho de Sistemas Computacionais e de Comunicação*. SBC, 2019.

[98] VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L.; GOMEZ, A. N.; KAISER, Ł.; POLOSUKHIN, I. **Attention is all you need**. *Advances in neural information processing systems*, 30, 2017.

[99] VINYALS, O.; FORTUNATO, M.; JAITLY, N. **Pointer networks**. *Advances in neural information processing systems*, 28, 2015.

[100] WAGNER FILHO, J. A.; WILKENS, R.; IDIART, M.; VILLAVICENCIO, A. **The brwac corpus: a new open resource for brazilian portuguese**. In: *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

[101] WEI, Q.; CHEN, T.; XU, R.; HE, Y.; GUI, L. **Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks**. *Database*, 2016:baw140, 2016.

[102] YADAV, V.; BETHARD, S. **A survey on recent advances in named entity recognition from deep learning models**. *arXiv preprint arXiv:1910.11470*, 2019.

[103] YE, J.; WU, Z.; FENG, J.; YU, T.; KONG, L. **Compositional exemplars for in-context learning**. In: *International Conference on Machine Learning*, p. 39818–39833. PMLR, 2023.

[104] YUTAO, F.; JIPENG, Q.; YUN, L.; YUNHAO, Y.; ZHU, Y. **Sentence simplification via large language models**. *arXiv preprint arXiv: 2302.11957 v1*, 2023.

[105] ZARATIANA, U.; TOMEH, N.; HOLAT, P.; CHARNOIS, T. **Gliner: Generalist model for named entity recognition using bidirectional transformer**. *arXiv preprint arXiv:2311.08526*, 2023.

[106] ZHENG, S.; WANG, F.; BAO, H.; HAO, Y.; ZHOU, P.; XU, B. **Joint extraction of entities and relations based on a novel tagging scheme**. *arXiv preprint arXiv:1706.05075*, 2017.

[107] ZHOU, G.; SU, J. **Named entity recognition using an hmm-based chunk tagger**. In: *Proceedings of the 40th annual meeting of the association for computational linguistics*, p. 473–480, 2002.

[108] ZHOU, P.; ZHENG, S.; XU, J.; QI, Z.; BAO, H.; XU, B. **Joint extraction of multiple relations and entities by using a hybrid neural network**. In: *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data: 16th China National Conference, CCL 2017, and 5th International Symposium, NLP-NABD 2017, Nanjing, China, October 13-15, 2017, Proceedings 16*, p. 135–146. Springer, 2017.

[109] ZHOU, W.; ZHANG, S.; GU, Y.; CHEN, M.; POON, H. **Universalner: Targeted distillation from large language models for open named entity recognition**. *arXiv preprint arXiv:2308.03279*, 2023.

[110] ZITOUNI, I. **Natural language processing of semitic languages**. Springer, 2014.

APÊNDICE A

Exemplo de *Prompt* no cenário *Zero-Shot*

System: Você é um algoritmo especialista em reconhecimento de entidades nomeadas em textos de editais de licitação.

A tarefa de Reconhecimento de Entidades Nomeadas tem como objetivo identificar e classificar entidades presentes em textos.

E uma entidade nomeada é uma expressão que pode ser extraída de um texto e é essencial para compreender um contexto.

Dado o conjunto de tipos de entidades:
[
'objeto_licitacao',
'valor_objeto',
'data_sessao',
'modalidade_licitacao',
'numero_exercicio',
'criterio_julgamento'
]
Extraía as entidades nomeadas no texto abaixo, e as classifique em algum dos tipos da lista.

Responda conforme o seguinte formato:
['entidade_nomeada_1': 'objeto_licitacao','entidade_nomeada_2': 'valor_objeto']

Os textos não necessariamente contêm entidades, portanto não adicione à lista os tipos que não têm entidades no texto.

Os textos podem ter mais de uma entidade de mesmo tipo, adicione todas entidades encontradas, com o tipos correspondentes, à lista na sua resposta.

Human: Texto: 2.1 NATUREZA DOS SERVIÇOS E FORMA DE SUA EXECUÇÃO : O objeto deste contrato é a PRESTAÇÃO DE SERVIÇOS DE ENGENHARIA PARA _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ (objeto da licitação) _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ , sob o regime de empreitada global . Deverão ser obedecidos os projetos , plantas , especificações , cronograma físico-financeiro e observações técnicas fornecidas pelo Município de Santa Rita do Novo Destino , que fazem parte integrante deste contrato .

APÊNDICE B

Exemplo de *Prompt* no cenário *Few-Shot*

System: Você é um algoritmo especialista em reconhecimento de entidades nomeadas em textos de editais de licitação.

A tarefa de Reconhecimento de Entidades Nomeadas tem como objetivo identificar e classificar entidades presentes em textos.

E uma entidade nomeada é uma expressão que pode ser extraída de um texto e é essencial para compreender um contexto.

Dado o conjunto de tipos de entidades:
[
'objeto_licitacao',
'valor_objeto',
'data_sessao',
'modalidade_licitacao',
'numero_exercicio',
'criterio_julgamento'
]
Extraía as entidades nomeadas no texto abaixo, e as classifique em algum dos tipos da lista.

Responda conforme o seguinte formato:
['entidade_nomeada_1': 'objeto_licitacao','entidade_nomeada_2': 'valor_objeto']

Os textos não necessariamente contêm entidades, portanto não adicione à lista os tipos que não têm entidades no texto.

Os textos podem ter mais de uma entidade do mesmo tipo, adicione todas as entidades encontradas, com seus tipos correspondentes, à lista na sua resposta.

Abaixo, alguns exemplos de sentenças com extração de entidades:
Texto: 1.1 – Constitui objeto do presente contrato é _ , conforme condições e especificações estabelecidas neste instrumento contratual e Edital da Tomada de Preços nº 01 2021 e seus Anexos : memorial descritivo , especificações técnicas , planilha orçamentária , composição de custos , cronograma físico – financeiro e projetos , nos termos do que dispõe a Lei 8.666 93 .

Resposta: [Tomada de Preços: modalidade_licitacao]

Texto: 2.1 – Constitui objeto da presente licitação a Contratação de Empresa Especializada para Prestação de Serviços de Engenharia no regime de empreitada global para Construção de um galpão em estrutura metálica para sediar as instalações da Feira Municipal , de acordo com projeto , memorial descritivo , cronograma físico financeiro e planilha orçamentária , composição BDI , que fazer parte do edital .

Resposta: [Contratação de Empresa Especializada para Prestação de Serviços de Engenharia no regime de empreitada global para Construção de um galpão em estrutura metálica para sediar as instalações da Feira Municipal: objeto_licitacao]