Proposta recebida em Setembro 2024 e aceite para publicação em Janeiro 2025.

Aprimorando o Reconhecimento de Entidades Nomeadas em Textos Literários em Português com Modelos Adaptativos

Enhancing Named Entity Recognition in Portuguese Literary Texts with Adaptive Models

Mariana O. Silva
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

Mirella M. Moro
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais

Resumo

Neste trabalho, investigamos estratégias de pré-treinamento para aprimorar o Reconhecimento de Entidades Nomeadas (REN) em textos literários em português. Introduzimos dois modelos adaptativos ao domínio, LitBERT-CRF e LitBERTimbau, construídos sobre modelos de linguagem de domínio geral. Avaliamos o aprendizado por transferência entre domínios em conjunto com um *baseline* de domínio geral (BERT-CRF). Nossas análises destacam a eficácia dessas estratégias e suas implicações para tarefas de REN literário. Resultados experimentais revelam que os modelos adaptados e ajustados ao domínio literário superam o *baseline*, alcançando uma pontuação F1 maior que 75% em um cenário de avaliação estrita e que 80% em um cenário parcial.

Palavras chave

reconhecimento de entidades nomeadas, pré-treinamento adaptativo, literatura em português

Abstract

We investigate pre-training strategies to enhance Named Entity Recognition (NER) in Portuguese literary texts. We introduce two domain-adaptive models, LitBERT-CRF and LitBERTimbau, built on general-domain language models. We also evaluate transfer learning across domains alongside a general-domain baseline (BERT-CRF). Overall, our findings highlight the efficiency of our strategies and their implications for literary NER tasks. Furthermore, experimental results reveal the adapted and domain-specific models outperform the generic baseline with an F1 score of over 75% in a strict evaluation scenario and over 80% in a partial scenario.

Keywords

named entity recognition, adaptive pre-training, literature in portuguese

1. Introdução

A literatura, considerada reflexo da cultura e da história, é rica em personagens diversos, lugares e alusões culturais. O Reconhecimento de Entidades Nomeadas (REM), como uma tarefa de Processamento de Linguagem Natural (PLN), possui uma importância profunda neste domínio (Claro et al., 2023; Vieira & Silva, 2023). Ao categorizar elementos literários essenciais, como nomes de personagens e locais, é possível analisar narrativas complexas, identificar padrões, acompanhar o desenvolvimento dos personagens e explorar o contexto sociocultural nas obras literárias.

Modelos de linguagem baseados em BERT (*Bidirectional Encoder Representations from Transformers*) (Devlin et al., 2019) têm se destacado em tarefas de REN. Combinados com redes recorrentes bidirecionais, mecanismos de atenção ou *Conditional Random Fields* (CRF), esses modelos demonstraram sua capacidade em capturar o contexto e as relações, tornando-os particularmente adequados para a complexidade das entidades literárias (Emelyanov & Artemova, 2019; Rodrigues et al., 2022). No entanto, o potencial de tais modelos depende da disponibilidade de dados rotulados para ajuste fino, um recurso que ainda é um tanto escasso no contexto de REN em Literatura escrita em português.

Textos literários, especialmente escritos em português, apresentam desafios únicos devido às complexidades da língua e à riqueza das alusões culturais (Santos et al., 2022). Além disso, a anotação de entidades nomeadas em textos literários apresenta seus próprios desafios únicos devido ao uso frequentemente metafórico ou simbólico de nomes, ao contexto histórico e à ambiguidade inerente aos papéis de personagens (Bamman et al., 2019; de Oliveira et al., 2022).

Para mitigar a questão da escassez de dados rotulados, modelos de linguagem, pré-treinados em grandes corpora não rotulados e ajustados

DOI: 10.21814/lm.17.1.443

em conjuntos de dados rotulados, revolucionaram o cenário das tarefas de PLN (Qiu et al., 2020; Raffel et al., 2020; Gururangan et al., 2020; Boukkouri et al., 2022). No entanto, esses modelos geralmente se originam de corpora genéricos e de domínio geral, como a Wikipédia. Em domínios especializados como a Literatura, essa abordagem pode ser subótima devido às profundas disparidades na terminologia específica do domínio, nas complexidades contextuais e nas nuances linguísticas (Bamman et al., 2019).

À luz desses desafios e lacunas de pesquisa, o objetivo deste trabalho é investigar e comparar diferentes estratégias de pré-treinamento para modelos baseados em BERT, destinados à tarefa de REN em Literatura escrita em português. Este trabalho estende o artigo apresentado no 16º *International Conference on Computational Processing of Portuguese* – PROPOR 2024 (Silva & Moro, 2024a). Como novas contribuições, uma análise detalhada dos erros cometidos pelos modelos ajustados é realizada, discutindo as causas e implicações. As principais contribuições deste artigo são as seguintes:

- Desenvolvimento de dois modelos adaptativos ao domínio literário, LitBERT-CRF e LitBERTimbau, que incorporam dados literários específicos durante o pré-treinamento (Seção 4) e ajuste fino (Seção 5);
- Comparação do desempenho desses modelos com um *baseline* de domínio geral, destacando os benefícios e *trade-offs* das estratégias de pré-treinamento (Seção 6); e
- Análise abrangente dos erros de classificação das entidades nomeadas, incluindo padrões de erros comuns e possíveis soluções para melhorar o desempenho dos modelos (Seção 7).

2. Estratégias de Pré-treinamento

Esta seção elabora conceitos sobre estratégias de pré-treinamento necessários para melhor compreender as contribuições deste artigo, bem como a discussão sobre trabalhos relacionados (Seção 3).

Em domínios altamente especializados, os modelos de linguagem genéricos podem ser subótimos devido às diferenças específicas do domínio. Consequentemente, existem pesquisas sobre estratégias de pré-treinamento para criar modelos de linguagem específicos para determinados contextos. Exemplos notáveis incluem aplicações clínicas (Lee et al., 2020), pesquisa científica (Beltagy et al., 2019), análise financeira (Liu et al., 2020) e indústria de petróleo e gás (Rodrigues et al., 2022; Freitas et al., 2023).

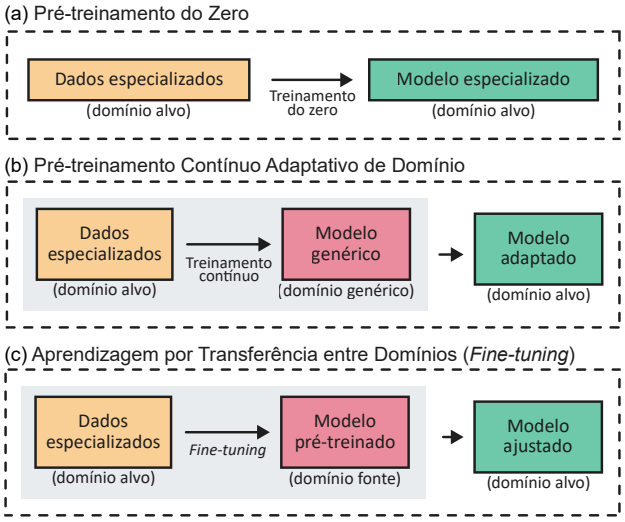

Figura 1: Estratégias de pré-treinamento.

No geral, o pré-treinamento específico de domínio requer dados do próprio domínio e pode ser realizado através de duas estratégias principais: começando do zero com um modelo treinado inteiramente em dados específicos do domínio, ou realizando um pré-treinamento contínuo de um modelo de linguagem genérico existente (Lamproudis & Henriksson, 2022). A primeira estratégia envolve treinar um modelo completamente novo, inicializado com pesos aleatórios, em um corpus substancial de dados do domínio. Embora esta abordagem exija uma quantidade considerável de dados, além de recursos computacionais e tempo substanciais, ela pode resultar em um modelo altamente especializado para o domínio alvo (Figura 1(a)).

Por outro lado, a abordagem adaptativa de domínio (Figura 1(b)) envolve a continuação do pré-treinamento de um modelo de linguagem genérico utilizando dados de texto não rotulados específicos do domínio (Qiu et al., 2020; Rodríguez et al., 2023). Tal estratégia é geralmente mais eficiente em termos de recursos e mais rápida do que o pré-treinamento do zero, sendo vantajosa tanto em cenários com alta quanto com baixa disponibilidade de recursos (Gururangan et al., 2020; Singhal et al., 2023).

Além do pré-treinamento específico de domínio, outra abordagem é a aprendizagem por transferência entre domínios, também conhecida como *fine-tuning* (Figura 1(c)). Tal estratégia é eficaz quando há dados anotados limitados no domínio alvo, mas um modelo já bem pré-treinado em um domínio fonte (Raffel et al., 2020). A aprendizagem por transferência aproveita o conhecimento adquirido no domínio fonte e o aplica ao domínio alvo, resultando em modelos que geralmente apresentam melhor

desempenho e convergência mais rápida durante o treinamento do que aqueles que começam do zero (Mou et al., 2016; Zhuang et al., 2021).

No geral, a escolha da estratégia de pré-treinamento depende das características específicas do domínio e da disponibilidade de recursos. Modelos de linguagem especializados são especialmente pertinentes em áreas onde a terminologia e os padrões linguísticos são altamente especializados, como na medicina, ciência, finanças e literatura. Ao adaptar os modelos para atender às necessidades desses domínios, é possível melhorar a precisão e a eficácia das aplicações de linguagem natural, tornando-as mais relevantes e úteis para usuários e pesquisadores.

Entre as diferentes estratégias de pré-treinamento existentes, neste estudo, são avaliadas duas: (i) pré-treinamento específico de domínio e (ii) aprendizado por transferência entre domínios. Tais abordagens foram escolhidas por sua capacidade de adaptar modelos de linguagem a contextos literários específicos de forma eficaz, diante da escassez de dados anotados disponíveis em literatura em português. Assim, este estudo busca não apenas analisar o desempenho dessas estratégias, mas também contribuir para o avanço do reconhecimento de entidades nomeadas em textos literários.

3. Trabalhos Relacionados

As estratégias de pré-treinamento desempenham um papel crucial em melhorar a eficácia de modelos de linguagem, especialmente em tarefas especializadas como análise de sentimentos, tradução automática e o reconhecimento de entidades nomeadas (REN). A tarefa de REN, que envolve a extração de entidades como nomes de pessoas, locais e organizações do texto (Claro et al., 2023; Vieira & Silva, 2023), pode ser significativamente aprimorada por meio do uso de modelos de linguagem adaptados a contextos específicos (Rodríguez et al., 2023; Docío et al., 2023).

Tais modelos são projetados para capturar nuances e terminologias de um domínio, resultando em um desempenho superior na identificação e classificação dessas entidades em comparação com modelos genéricos. No entanto, em relação às obras literárias, especialmente as escritas em português, os modelos de REN enfrentam desafios únicos. A literatura muitas vezes reflete cultura, história e personagens diversos, produzindo um conteúdo rico, mas complexo.

Portanto, identificar e extrair efetivamente entidades nesses contextos requer estratégias de pré-treinamento adaptativas de domínio que integrem as especificidades culturais e estilísticas das obras. Para isso, é fundamental desenvolver corpora anotados que reflitam a diversidade linguística de obras literárias. Por exemplo, Bamman et al. (2019) apresentam um corpus de 100 textos literários em inglês, anotados para categorias de entidades nomeadas. Os resultados do estudo mostram que modelos adaptados ao domínio literário melhoraram o F-score em mais de 20 pontos absolutos (de 45,7 para 68,3).

De forma mais ampla, Frontini et al. (2020) apresentam a Coleção de Textos Literários Europeus (ELTeC), um recurso multilíngue e de código aberto. O corpus ELTeC é composto por uma diversidade de textos literários europeus que possibilitam a avaliação e o desenvolvimento de ferramentas de REN adaptadas a diferentes idiomas e contextos culturais. Em sua versão atual, o corpus contém textos literários anotados em mais de dez línguas diferentes, incluindo o português. Santos et al. (2020) relatam a preparação da anotação do ELTeC-por e apresentam um sistema de reconhecimento de entidades nomeadas, o PALAVRAS-NER (Bick, 2000), que é primordialmente baseado em regras.

Diante de tais trabalhos, este artigo se diferencia ao focar no desenvolvimento e na avaliação de modelos de REN adaptados ao domínio literário em português. Enquanto estudos anteriores exploram corpora literários em inglês e múltiplos idiomas europeus, o presente estudo busca não apenas melhorar o desempenho dos modelos de REN em textos literários em português, mas também contribuir para a compreensão mais ampla de como os modelos de linguagem podem ser adaptados para diferentes contextos linguísticos, oferecendo ideias para futuras pesquisas na área de processamento de linguagem natural.

4. Pré-treinamento

Esta seção apresenta as metodologias utilizadas para o pré-treinamento dos modelos de linguagem propostos. Na Seção 4.1, é descrita a construção do corpus utilizado para o pré-treinamento, enquanto as Seções 4.2 e 4.3 detalham as configurações experimentais e os modelos desenvolvidos, respectivamente.

4.1. Corpus

Para criar o corpus de pré-treinamento, o conjunto de dados *PPORTAL* (Silva et al., 2021, 2022)[1] foi utilizado para selecionar um subcon-

[1] https://doi.org/10.5281/zenodo.5178063

Obra	Autoria	Língua	Ano	Gênero	Movimento	#Tokens
A Cidade e as Serras	José Maria Eça de Queirós	pt	1901	Romance	Realismo	83.280
A Escrava Isaura	Bernardo Guimarães	pt-br	1875	Romance	Romantismo	62.358
A Relíquia	José Maria Eça de Queirós	pt	1887	Romance	Realismo	103.808
Cinco Minutos	José de Alencar	pt-br	1876	Romance	Romantismo	17.041
Clepsidra	Camilo Pessanha	pt	1920	Poesia	Simbolismo	6.033
Do Livro do Desassossego	Fernando Pessoa	pt	1982	Prosa poética	Modernismo	7.552
Iaiá Garcia	Machado de Assis	pt-br	1878	Romance	Romantismo	68.192
Noite na Taverna	Álvares de Azevedo	pt-br	1855	Contos	Romantismo	22.380
O Guarani	José de Alencar	pt-br	1857	Romance	Romantismo	124.540
Senhora	José de Alencar	pt-br	1875	Romance	Realismo	88.604
					Total	583.788

Tabela 1: Visão geral do corpus, usando no pré-treinamento.

junto de obras. O *PPORTAL* é um extenso repositório de metadados que contém mais de 80.000 obras literárias de domínio público na língua portuguesa, predominantemente derivadas do Brasil e Portugal. O conjunto de dados também fornece links para download de mais de 9500 obras completas, facilitando o acesso ao conteúdo textual necessário para o pré-treinamento.

O subconjunto selecionado para este estudo inclui 583.788 tokens de dez obras literárias de domínio público. Para garantir a qualidade do corpus, cada texto passou por pré-processamento, que inclui a remoção de caracteres especiais (exceto hífens e sinais de pontuação relevantes em contextos literários) e referências a e-mails e sites. A Tabela 1 resume as principais características do corpus final.

O corpus abrange uma variedade de gêneros literários, incluindo romances, poesias e contos, representando tanto o Romantismo quanto o Realismo, além de movimentos literários mais contemporâneos como o Modernismo. Além disso, a variação temporal das obras, que abrange desde o século XIX até o século XX, proporciona um panorama abrangente das transformações linguísticas e estilísticas ao longo do tempo.

4.2. Configuração

Todas as sessões de pré-treinamento utilizam a tarefa de *Masked Language Modeling* (MLM). Nessa tarefa, uma porcentagem predeterminada de palavras dentro de uma sequência (especificamente 15%) é deliberadamente mascarada, e o principal objetivo do modelo é prever essas palavras. Esse processo não apenas aprimora a compreensão do modelo sobre as relações contextuais da linguagem, mas também melhora sua capacidade de gerar texto coeso.

A duração máxima de três épocas foi estabelecida para equilibrar recursos computacionais e limitações de tempo, garantindo que os modelos pudessem se beneficiar de várias iterações sem comprometer a viabilidade do treinamento. Foram utilizados os mesmos hiperparâmetros para todos os modelos pré-treinados, conforme detalhado na Tabela 2. Além disso, em vez de avaliar diretamente a tarefa de pré-treinamento, cada *checkpoint* salvo é analisado em termos de desempenho na tarefa final de reconhecimento de entidades nomeadas (Seção 6).

Hiperparâmetro	Valor
Learning rate	5×10^{-5}
Batch size	16
Comprimento máximo	512
Épocas	3
Probabilidade de MLM	15%

Tabela 2: Configuração de hiperparâmetros usados durante o pré-treinamento.

4.3. Modelos Pré-treinados

Neste artigo, dois novos modelos de linguagem foram desenvolvidos para o reconhecimento de entidades nomeadas. Ambos os modelos são pré-treinados utilizando a tarefa de MLM em nosso subconjunto de dados, incorporando características específicas do domínio para melhorar a identificação e o reconhecimento de entidades em textos literários escritos em português. A seguir, cada modelo é descrito brevemente.[2]

LitBERTimbau. Este modelo é baseado no BERTimbau-Base (Souza et al., 2020), que foi inicialmente pré-treinado sobre o brWAC (Brazilian Web Corpus) (Filho et al., 2018). O BERTimbau-Base possui uma arquitetura de Transformer, contendo 110M parâmetros e for-

[2]Ambos modelos pré-treinados estão disponibilizados em: https://huggingface.co/marianaossilva.

nece uma sólida base na compreensão da língua portuguesa e no conhecimento linguístico geral, permitindo que LitBERTimbau se beneficie de suas capacidades em um contexto literário.

LitBERT-CRF. Este modelo aproveita a arquitetura BERT-CRF (Souza et al., 2019), que combina o modelo BERT-Base com *Conditional Random Fields* (CRF) para aprimorar o reconhecimento de entidades nomeadas. O checkpoint utilizado do modelo BERT-CRF também foi pré-treinado no corpus brWaC (Filho et al., 2018) e posteriormente ajustado com o corpus First HAREM (Santos et al., 2006), que contém entidades nomeadas rotuladas em português. Esta combinação permite que o LitBERT-CRF obtenha melhores resultados em tarefas de REN devido à sua arquitetura especializada.

5. Ajuste Fino (*Fine-tuning*)

O ajuste fino (ou *fine-tuning*), que corresponde ao aprendizado por transferência entre domínios, é um processo em que um modelo previamente treinado em um domínio fonte é ajustado para realizar tarefas em um domínio alvo. Esta seção apresenta as metodologias utilizadas para o ajuste fino de três modelos pré-treinados. Na Seção 5.1, é apresentado o corpus utilizado para o ajuste fino, enquanto as Seções 5.2 e 5.3 detalham as configurações experimentais e os modelos ajustados, respectivamente.

5.1. Corpus

Para ajustar e avaliar os modelos pré-treinados em uma tarefa final, foi considerado o corpus *PPORTAL_ner*, que foi manualmente anotado para entidades nomeadas (Silva & Moro, 2024b). O *PPORTAL_ner* também é proveniente do conjunto de dados *PPORTAL* e contém um conjunto de 25 obras literárias individuais.[3] Todos os textos incluídos no *PPORTAL_ner* foram publicados antes de 1953, atendendo aos critérios atuais de domínio público no Brasil, com a maioria das obras datando entre 1554 e 1938. No total, o corpus possui 125.059 tokens, 5.418 sentenças e 5.266 entidades anotadas.

O processo de anotação foi realizado por um único anotador e seguiu um protocolo em duas etapas: pré-anotação inicial utilizando o modelo spaCy *pt_core_news_lg* e posterior correção e refinamento com a ferramenta de anotação Prodigy.[4] Embora reconhecendo a limitação de um único anotador, o esforço foi feito para garantir a consistência e precisão ao longo do processo de anotação. O corpus final inclui anotações de cinco classes de entidades:

- PER (Pessoa)
- LOC (Localização)
- GPE (Entidade geopolítica)
- ORG (Organização)
- DATE (Data)

Classe	Frequência (%)	Exemplos
PER	3.609 (68,53%)	“Capitu”, “o estrangeiro”
LOC	1.126 (21,38%)	“a aldeia”, “a cidade”
GPE	315 (5,98%)	“Brasil”, “Lisboa”
ORG	115 (2,18%)	“a polícia”, “a Igreja”
DATE	101 (1,92%)	“século XVIII”, “1847”

Tabela 3: Distribuição das classes de entidades.

A Tabela 3 apresenta uma análise detalhada da frequência de cada classe de entidade, expressa como porcentagem do total de entidades anotadas, juntamente com exemplos ilustrativos que destacam o conteúdo do corpus. Note que as regras de anotação do *PPORTAL_ner* foram projetadas para capturar a complexidade dos textos literários, incluindo tanto nomes próprios quanto expressões descritivas usadas para referenciar entidades (e.g., “o estrangeiro”, “a aldeia”). Essa abordagem difere de modelos tradicionais, como o HAREM (Santos et al., 2006), que foca em nomes próprios estritos, mas se alinha à tarefa de detecção de entidades definida no ACE (Doddington et al., 2004), que admite nomes, pronomes e expressões descritivas.

Além disso, o processo de anotação do corpus *PPORTAL_ner* adota uma definição abrangente para a classe PER (Pessoa), considerando como tal qualquer indivíduo que desempenhe um papel como personagem dentro da narrativa, independentemente de sua natureza biológica. Essa definição abrange tanto personagens humanos quanto não humanos, incluindo animais personificados, figuras mitológicas ou entidades ficcionais que possuam características e ações que os qualifiquem como participantes ativos na trama. Por exemplo, em uma fábula onde um lobo age e se comunica como um ser humano, o lobo seria anotado como PER.

5.2. Configuração

Durante o processo de ajuste fino, os três modelos pré-treinados são ajustados por um número fixo

[3] Note que o corpus usado durante o pré-treinamento compreende obras literárias diferentes das 25 obras selecionadas para o ajuste fino.

[4] https://prodi.gy/

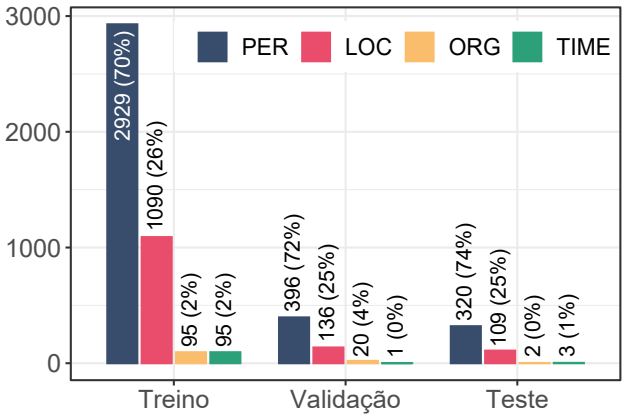

Figura 2: Distribuição das entidades nos conjuntos de treinamento, validação e teste.

de dez épocas. Essa duração foi selecionada para equilibrar a necessidade de recursos computacionais e garantir que os modelos se ajustem adequadamente ao novo domínio. Não é realizada uma busca extensiva de hiperparâmetros, pois o foco principal é avaliar a eficácia das diferentes estratégias na criação de modelos de linguagem literária (em vez de otimizar o desempenho em relação a uma tarefa específica).

As sessões de ajuste fino envolvem monitoramento contínuo para assegurar a convergência da perda do conjunto de validação. A convergência é avaliada com base na estabilidade do desempenho em termos de perda e precisão, indicando que os modelos estão aprendendo e se adaptando às características específicas do corpus literário. Além disso, utiliza-se um conjunto de validação separado para evitar o *overfitting*, garantindo que os modelos mantenham uma capacidade de generalização em dados não vistos.

O corpus anotado é, portanto, dividido em conjuntos de treinamento, validação e teste. Essa divisão aloca 80% das sentenças para o conjunto de treinamento (5.572 sentenças), 10% para o conjunto de validação (696 sentenças) e os 10% restantes para o conjunto de teste (697 sentenças). A Figura 2 apresenta a distribuição das entidades em cada classe dentro dos conjuntos de treinamento, validação e teste, mostrando que o corpus possui um desbalanceamento significativo entre as classes de entidades. Esse desbalanceamento pode influenciar o desempenho dos modelos, levando-os a ter maior dificuldade em identificar classes de entidades menos representadas.[5]

[5]Essa questão de desbalanceamento será abordada em trabalhos futuros, explorando estratégias de reamostragem ou ajuste de pesos para melhorar a performance do modelo em classes sub-representadas.

5.3. Modelos Ajustados

Como mencionado anteriormente, três modelos pré-treinados passam pelo processo de ajuste fino: (i) LitBERTimbau, (ii) LitBERT-CRF e (iii) BERT-CRF. Os dois primeiros modelos são inicialmente pré-treinados com um corpus literário específico (veja Seção 4), utilizando a estratégia de pré-treinamento de domínio. O terceiro modelo, BERT-CRF, é um modelo de domínio geral ajustado para melhorar sua capacidade de reconhecimento de entidades nomeadas em textos literários através do aprendizado por transferência entre domínios.

O BERT-CRF ajustado usando o corpus literário específico (*PPORTAL_ner*) é referenciado aqui como FT BERT-CRF. Além do modelo FT BERT-CRF, utilizamos como *baseline* o modelo BERT-CRF puro, ajustado originalmente utilizando o corpus genérico HAREM (Santos et al., 2006), que possui duas versões distintas. A primeira versão contém um conjunto de dez classes de entidades nomeadas, enquanto a segunda versão ("selective") foca em cinco classes: PER, ORG, LOC, VALUE e TIME.[6] Para garantir a consistência entre os corpora, as entidades GPE do corpus literário são reclassificadas como LOC e as entidades DATE como TIME. A Tabela 4 resume as principais características de cada modelo considerado.

Como foi dito anteriormente, o *PPORTAL_ner* adota uma conceituação mais ampla de entidades nomeadas, incluindo nomes próprios e expressões descritivas. Essa abordagem contrasta com as diretivas do HAREM, que se baseiam em uma definição mais restritiva, centrada em designadores rígidos, como nomes próprios. Tal diferença de escopo dificulta a comparação dos modelos, uma vez que o desempenho de modelos ajustados com base no *PPORTAL_ner* pode refletir sua capacidade de identificar construções semânticas mais complexas, enquanto os modelos baseados no HAREM são avaliados em um conjunto mais delimitado de entidades.

Modelo	**Vocab**	**C1**	**C2**
BERT-CRF	Genérico	Genérico	Genérico
FT BERT-CRF	Genérico	Genérico	Literário
LitBERT-CRF	Genérico	Literário	Literário
LitBERTimbau	Genérico	Literário	Literário

C1: Corpus de pré-treinamento
C2: Corpus de ajuste fino

Tabela 4: Visão geral dos modelos ajustados.

[6]PER (Pessoa), ORG (Organização), LOC (Localização), VALUE (Valor) e TIME (Tempo).

6. Avaliação Experimental

Esta seção descreve a avaliação experimental para analisar as diferentes estratégias de pré-treinamento. Primeiramente, as métricas de avaliação são apresentadas na Seção 6.1 e, em seguida, os resultados são discutidos na Seção 6.2.

6.1. Métricas de Avaliação

Ao avaliar modelos de REN, é prática comum relatar métricas no nível individual do token. No entanto, essa abordagem pode não ser sempre a mais abrangente, especialmente considerando que uma entidade nomeada pode abranger múltiplos tokens. Para fornecer uma avaliação mais precisa, é essencial considerar a acurácia de entidades completas.

Portanto, neste estudo, é adotado o esquema de avaliação definido pela *SemEval 2013 - 9.1 task* (Segura-Bedmar et al., 2013), que vai além de um esquema simples baseado em token/tag. Tal esquema considera diferentes cenários, verificando se todos os tokens pertencentes a uma entidade nomeada são corretamente classificados e se o tipo de entidade correto foi atribuído.

Tipo	**Descrição**
Correto (C)	Entidades reais e preditas são iguais
Incorreto (I)	Entidades reais e preditas não coincidem
Parcial (P)	Entidades reais e preditas são similares
Ausente (M)	Entidade real que não foi predita
Espúrio (S)	Entidade predita que não existe

Tabela 5: Tipos de erros na avaliação.

Cenário	**Descrição**
Estrito	Correspondência exata de limite e tipo
Tipo	Atribuição correta do tipo de entidade, independentemente dos limites exatos
Parcial	Correspondência parcial dos limites, independentemente do tipo de entidade
Exato	Correspondência exata dos limites, independentemente do tipo de entidade

Tabela 6: Cenários de avaliação.

Dentro de tal esquema de avaliação, cinco tipos de erros são considerados para capturar diferentes aspectos do desempenho dos modelos: Correto (C), Incorreto (I), Parcial (P), Ausente (M) e Espúrio (S). A Tabela 5 descreve cada tipo de erro. Além disso, quatro cenários distintos de avaliação são considerados, cada um avaliando o desempenho dos modelos de maneiras diferentes: Estrito, Tipo, Parcial e Exato. A Tabela 6 descreve esses cenários de avaliação.

Para avaliação automatizada, os erros são calculados com base na correspondência dos limites, especificamente avaliando se há uma sobreposição entre as entidades verdadeiras e preditas. A sobreposição é determinada pela interseção entre os deslocamentos de início e fim das entidades verdadeiras e preditas. Por exemplo, se a entidade verdadeira abrange do terceiro ao sétimo token, e a entidade predita abrange do quinto ao nono token, a sobreposição incluiria os tokens 5, 6 e 7. Essa abordagem permite uma avaliação detalhada da correspondência parcial dos limites sem impor restrições rígidas de porcentagem.

6.2. Resultados Experimentais

A Tabela 7 resume os resultados da avaliação dos modelos, baseados nos cinco tipos de erros e em quatro cenários de análise. Note que a quantidade de entidades ausentes (M) e espúrias (S) permanece a mesma, independentemente dos cenários de avaliação, uma vez que o cálculo dessas métricas não se baseia nos limites das entidades, mas sim na presença ou ausência de entidades previstas em relação às anotações de referência. Isso indica que a dificuldade dos modelos em capturar todas as entidades corretas e em evitar a identificação de entidades que não existem são consistentes, independentemente do rigor dos critérios de avaliação aplicados.

Além disso, ao analisar os resultados, observa-se que todos os modelos enfrentam desafios semelhantes relacionados à identificação precisa do tipo de entidade. Essa questão se torna mais evidente no cenário **estrito**, onde a exigência de correspondência exata dos limites e tipos leva a uma taxa maior de erros incorretos (I). Em contraste, nos cenários **tipo** e **parcial**, onde a correspondência é mais flexível, a quantidade de entidades corretamente identificadas (C) tende a aumentar, sugerindo que os modelos têm uma capacidade melhorada para reconhecer entidades quando não são impostas restrições rígidas.

Ao comparar com o *baseline* (BERT-CRF), todos os três modelos ajustados apresentam um desempenho geral robusto. O modelo BERT-CRF puro demonstra uma capacidade relativamente baixa de capturar entidades, evidenciada pelo seu número notavelmente baixo de entidades corretas (C). Além disso, registra uma taxa considerável de entidades ausentes (M), sugerindo

Modelo	Estrito				Tipo				Parcial				Exato			
	C	*I*	*M*	*S*	*C*	*I*	*M*	*S*	*C*	*P*	*M*	*S*	*C*	*I*	*M*	*S*
BERT-CRF	119	3	312	29	120	2	312	29	121	1	312	29	121	1	312	29
FT BERT-CRF	335	35	65	65	362	8	65	65	341	29	65	65	341	29	65	65
LitBERT-CRF	336	31	67	62	357	10	67	62	344	23	67	62	344	23	67	62
LitBERTimbau	333	33	68	84	358	8	68	84	341	25	68	84	341	25	68	84

C: Correto — *I:* Incorreto — *M:* Ausente — *S:* Espúrio — *P:* Parcial

Tabela 7: Resultados da avaliação com base nos cinco tipos de erros e quatro cenários. O conjunto de teste usado para avaliar os modelos tem 434 entidades anotadas.

desafios na identificação de certas entidades nomeadas. Também apresenta algumas entidades espúrias (S), indicando uma tendência a classificar incorretamente entidades que não existem.

Em contraste, os modelos ajustados mostram uma melhoria significativa na identificação correta de entidades nomeadas (C) em comparação com o *baseline*. No entanto, ainda apresentam um número relativamente elevado de entidades classificadas como incorretas (I), sugerindo que esses modelos ainda cometem erros na classificação de algumas entidades. Embora tenham reduzido a taxa de entidades ausentes (M) em comparação com o *baseline*, evidenciando uma maior capacidade de reconhecimento de entidades nomeadas, ainda enfrentam dificuldades na identificação de algumas entidades específicas.

Além da análise dos tipos de erros, também são calculadas as métricas de precisão, revocação e F1-Score para cada cenário (Tabela 8). Aqui, a precisão é o percentual de entidades nomeadas corretamente identificadas pelo modelo, enquanto a revocação representa a capacidade do modelo de capturar o percentual de entidades nomeadas nas anotações de ouro com sucesso. Já o F1-Score combina precisão e revocação em uma única métrica, particularmente útil para avaliar o desempenho em tarefas desbalanceadas.

O cálculo de tais métricas é conduzido de duas maneiras distintas, dependendo se uma correspondência exata é considerada necessária (para os cenários **estrito** e **exato**) ou se uma correspondência parcial é aceitável (para os cenários **parcial** e **tipo**). No geral, como detalhado a seguir, tanto o aprendizado por transferência entre domínios quanto o pré-treinamento adaptativo ao domínio demonstram ser abordagens valiosas para criar modelos de linguagem adaptados a tarefa de REN no contexto literário.

Aprendizado por transferência entre domínios (*fine-tuning*). No geral, o modelo BERT-CRF ajustado (FT BERT-CRF) demonstra um desempenho competitivo. Esse modelo aproveita o conhecimento pré-treinado de um domínio geral para se adaptar ao domínio literário, capturando significativamente entidades na tarefa de REN. Para o cenário **estrito**, o modelo apresenta um F1-Score de 77% com um equilíbrio entre precisão e revocação. Esse desempenho equilibrado indica que o modelo se destaca em identificar corretamente as entidades e capturar uma proporção significativa das entidades nomeadas.

No cenário **exato**, que avalia a correspondência exata de limites independentemente do tipo de entidade, o modelo também apresenta um alto F1-Score (78%). Esse resultado destaca sua capacidade de capturar uma parte substancial das entidades nomeadas enquanto mantém uma correspondência precisa dos limites. Ao considerar cenários de correspondência de limites mais relaxados, como **tipo** e **parcial**, o FT BERT-CRF supera as expectativas com um F1-Score superior a 81%. Ou seja, o modelo pode capturar uma maior proporção de entidades nomeadas quando os limites não são exatos.

Em comparação com os outros dois modelos pré-treinados (LitBERT-CRF e LitBERTimbau), o aprendizado por transferência entre domínios mostra resultados fortes, especialmente em cenários onde a correspondência exata de limites não é necessária. Os resultados competitivos do modelo podem ser atribuídos à sua capacidade de aproveitar o amplo conhecimento linguístico e contextual em dados de domínio geral, acelerando assim sua transição para o domínio literário.

Pré-treinamento adaptativo ao domínio. No geral, o modelo LitBERT-CRF supera os outros modelos na maioria dos cenários de avaliação na Tabela 8. No entanto, seu desempenho está alinhado de perto com o modelo BERT-CRF ajustado. Mas, diferentemente do aprendizado por transferência entre domínios, que se adapta ao domínio literário aproveitando o conhecimento prévio de um domínio geral, o pré-treinamento adaptativo ao domínio incorpora diretamente dados específicos do domínio no processo de pré-

Modelo	Estrito			Tipo			Parcial			Exato		
	P	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*	*P*	*R*	*F1*
BERT-CRF	**0.788**	0.274	0.407	0.795	0.276	0.410	0.805	0.28	0.415	0.801	0.279	0.414
FT BERT-CRF	0.770	0.770	0.770	**0.832**	**0.832**	**0.832**	0.817	0.817	0.817	0.784	0.784	0.784
LitBERT-CRF	0.783	**0.774**	**0.779**	**0.832**	0.823	0.827	**0.829**	**0.819**	**0.824**	**0.802**	**0.793**	**0.797**
LitBERTimbau	0.740	0.767	0.753	0.796	0.825	0.810	0.786	0.815	0.800	0.758	0.786	0.771

P: Precisão — *R:* revocação — *F1:* F1-Score

Tabela 8: Resultados da avaliação dos modelos de NER em diferentes dados de treinamento. O melhor desempenho é destacado em negrito e o segundo melhor está sublinhado.

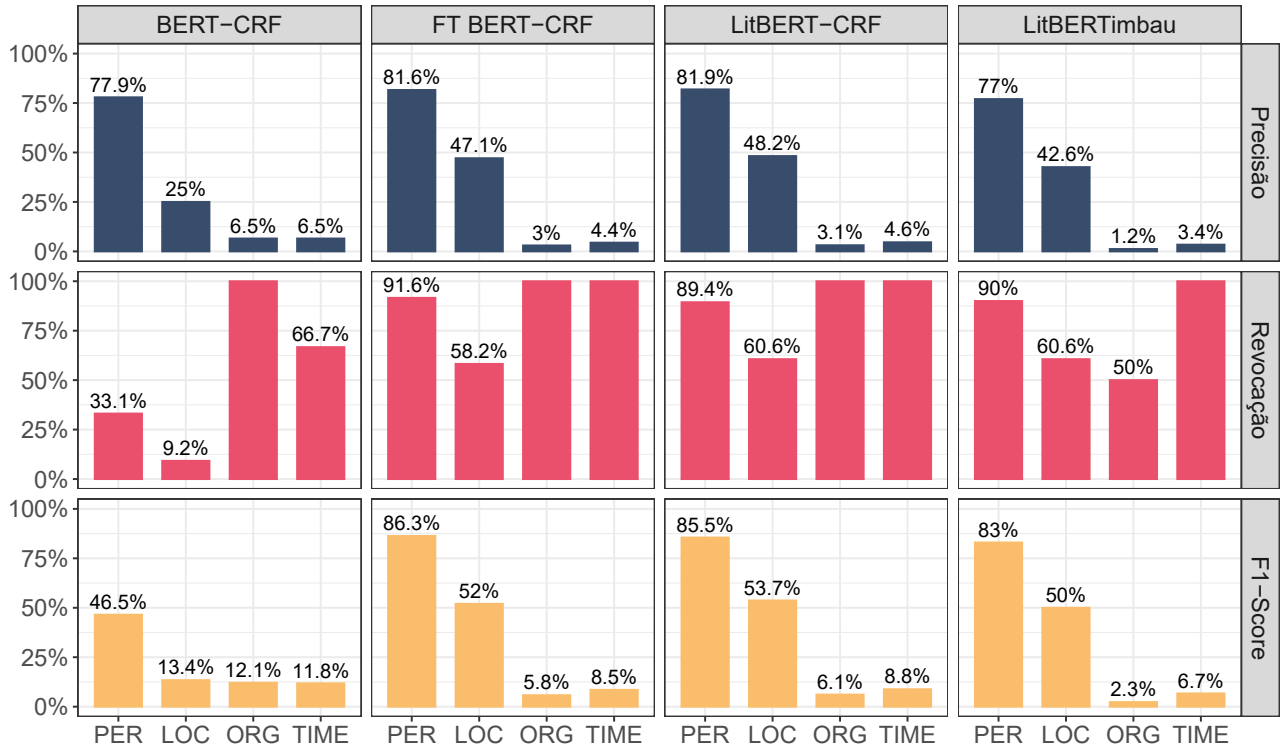

Figura 3: Métricas de avaliação para cada modelo, considerando o cenário **tipo**.

treinamento, potencialmente equipando o modelo com nuances linguísticas mais especializadas.

O forte desempenho do modelo LitBERT-CRF, particularmente no cenário **estrito**, enfatiza sua precisão em identificar precisamente entidades nomeadas, alcançando um F1-Score de 78%. Esse resultado sugere que o modelo não apenas identifica uma parte significativa das entidades, mas também as classifica com precisão. O desempenho consistentemente alto do modelo se estende a outros cenários, especialmente o **tipo** e **parcial**, com F1-Scores acima de 82%.

Em contraste, o modelo LitBERTimbau, que se baseia no modelo de domínio geral BERTimbau, apresenta pontuações de F1 competitivas, mas ligeiramente inferiores em todos os cenários de avaliação. Embora ele tenha um bom desempenho, fica um pouco aquém do nível de precisão do modelo LitBERT-CRF na identificação de entidades nomeadas. Essa discrepância pode ser atribuída a diferentes fatores.

Primeiro, o pré-treinamento inicial do BERTimbau no corpus brWaC pode fornecer uma base linguística ampla, mas talvez não seja tão adaptado às nuances literárias quanto o modelo LitBERT-CRF, que foi também pré-treinado no brWaC, mas ajustado utilizando o conjunto de dados HAREM I. Segundo, a diferença na complexidade dos modelos pode contribuir para os resultados observados. O LitBERT-CRF, ao combinar o BERT-base com *Conditional Random Fields* (CRF), pode melhorar a precisão no reconhecimento de entidades nomeadas, aproveitando a estrutura adicional do CRF para capturar dependências sequenciais mais complexas.

Tipo de entidade. A Figura 3 mostra métricas de avaliação ao nível da entidade para cada modelo, focando exclusivamente no cenário **tipo**. Nesse cenário específico, é permitida uma certa

Modelo	Ausente				Espúrio	Incorreto-Estrito				Incorreto-Tipo				Parcial			
	P	*L*	*O*	*T*	-	*P*	*L*	*O*	*T*	*P*	*L*	*O*	*T*	*P*	*L*	*O*	*T*
BERT-CRF	213	98	0	1	29	2	1	0	0	1	1	0	0	1	0	0	0
FT BERT-CRF	26	39	0	0	65	15	20	0	0	1	7	0	0	14	15	0	0
LitBERT-CRF	33	34	0	0	62	12	19	0	0	1	9	0	0	11	12	0	0
LitBERTimbau	30	38	0	0	84	13	18	2	0	2	5	1	0	11	13	1	0

P: PER — *L:* LOC — *O:* ORG — *T:* TIME

Tabela 9: Tipos de erros cometidos pelos modelos avaliados, pelo tipo de entidade.

sobreposição entre os limites das entidades verdadeiras e preditas, o que adiciona uma camada de flexibilidade à avaliação. Para complementar, a Tabela 9 apresenta os tipos de erros cometidos pelos modelos, por tipo de entidade.

Em comparação com o *baseline*, os três modelos ajustados mostram altas pontuações de avaliação para a classe de entidade PER (pessoa). Embora o LitBERT-CRF atinja uma precisão mais alta (82%), o modelo exibe uma taxa de revocação mais baixa, o que resulta em um F1-Score ligeiramente inferior (85,5%) em comparação com o modelo FT BERT-CRF (86,3%). De acordo com a Tabela 9, a revocação mais baixa do LitBERT-CRF pode ser atribuída ao maior número de erros da categoria *ausente*, onde entidades PER presentes no texto não foram detectadas pelo modelo.

O desempenho geral sólido na identificação e categorização correta de entidades PER em todos os modelos é esperado, pois essa classe de entidade é relativamente bem reconhecida pelo modelo *baseline* genérico. De fato, as entidades PER são frequentemente mais comuns e mais evidentes em textos literários, facilitando sua detecção em comparação com outras classes de entidades, tanto para modelos de linguagem genéricos quanto adaptativos ao domínio (Li et al., 2022).

No que diz respeito às outras classes de entidades, os modelos avaliados apresentam resultados de desempenho mais variados. Especificamente para a classe LOC (localização), os modelos específicos do domínio alcançam pontuações de F1 mais altas em comparação com o *baseline*. Esse resultado sugere que a incorporação de conhecimento específico do domínio (ou seja, dados literários) por meio de estratégias de pré-treinamento melhora significativamente a extração de entidades de localização em textos literários em português. De fato, de acordo com a Tabela 9, tais modelos apresentaram um número muito menor de entidades *ausentes* em comparação com o *baseline*.

No entanto, os modelos adaptados apresentaram um número muito maior de entidades LOC incorretas nos cenários **estrito** e **tipo**, e de entidades parciais. Isso indica que, embora os modelos tenham maior capacidade de identificar menções de localizações em textos literários, eles frequentemente erram na delimitação precisa dos limites das entidades ou na classificação exata do tipo da entidade. Esses erros podem ser atribuídos à complexidade intrínseca dos textos literários, exigindo um entendimento contextual mais profundo. Além disso, os erros parciais sugerem que os modelos são capazes de capturar fragmentos relevantes das entidades, mas enfrentam dificuldades em reconhecer toda a expressão como uma única unidade coesa.

Já ao avaliar as classes de entidades ORG (organização) e TIME (tempo), todos os modelos enfrentam desafios na identificação precisa. Apesar de alcançarem uma taxa de revocação relativamente alta, indicando sua capacidade de capturar uma parte substancial dessas entidades, a precisão dos modelos no reconhecimento de entidades ORG e TIME é notavelmente mais baixa. Isso sugere que, embora os modelos capturem muitas instâncias de organizações e expressões temporais, eles também geram numerosos falsos positivos, diminuindo a precisão.

A variabilidade no desempenho das entidades ORG e TIME pode ser atribuída à complexidade e diversidade de como as organizações e as expressões temporais são referenciadas em textos literários. Os autores costumam empregar maneiras criativas e dependentes do contexto para mencionar organizações e informações temporais, tornando a tarefa de REN um desafio para generalizar efetivamente (Cui & Joe, 2023).

Conceituação de entidades nomeadas. Um ponto relevante para entender os resultados obtidos está na diferença entre as conceituações de entidades nomeadas do *PPORTAL_ner* e do HAREM. O *PPORTAL_ner*, ao adotar uma abordagem mais ampla, que inclui nomes próprios e expressões descritivas, visa capturar a riqueza semântica típica de textos literários. Em contraste, o HAREM utiliza uma definição mais restrita, focada em designadores rígidos, como no-

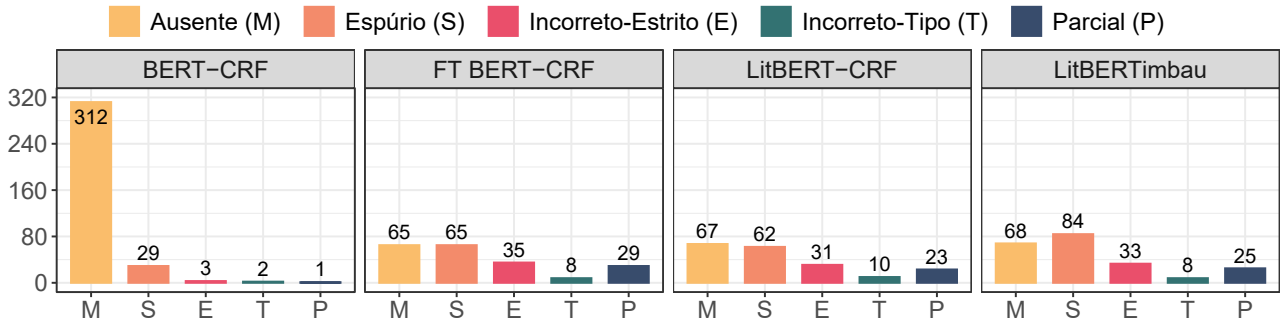

Figura 4: Distribuição de erros nos modelos avaliados.

mes próprios. Essa divergência conceitual pode ter afetado o desempenho dos modelos avaliados, especialmente no cenário **estrito**, onde a correspondência exata de limites e tipos é exigida. Modelos treinados com o *PPORTAL_ner* enfrentam desafios adicionais ao tentar acomodar categorias mais complexas e diversificadas, o que pode ter contribuído para o maior número de erros incorretos (I) observados, apesar das melhorias em outros cenários de avaliação.

7. Análise de Erro

Esta seção apresenta uma análise detalhada dos erros cometidos pelos modelos no reconhecimento de entidades nomeadas em textos literários em português. O objetivo é identificar as principais fontes de erro e explorar padrões recorrentes que possam orientar melhorias futuras. A análise está dividida em três sub-análises principais: entidades ausentes (Seção 7.1), entidades espúrias (Seção 7.2), e entidades incorretas e parciais (Seção 7.3). Por fim, são discutidas as implicações de tais erros e possíveis estratégias de mitigação (Seção 7.4).

A análise de erros é uma parte crucial na avaliação de modelos de reconhecimento de entidades nomeadas ao permitir identificar áreas específicas de melhoria e entender as limitações dos modelos. Nesta seção, são abordadas as principais categorias de erros observadas nos modelos avaliados, destacando os tipos de erros mais comuns, apresentando exemplos, discutindo suas possíveis causas e suas implicações para o desempenho geral dos modelos.

Como indicado na Tabela 5, os erros observados podem ser classificados em quatro categorias principais de entidades: ausentes (M), espúrias (S), incorretas (I) e parciais (P). A Figura 4 fornece uma visão geral da distribuição desses erros para cada modelo. A categoria *incorreta* é avaliada tanto nos cenários de avaliação **estrito** e **tipo**, enquanto a categoria *parcial* é avaliada apenas no cenário **parcial** (que corresponde à categoria *incorreta* no cenário **exato**).

Ausentes (M). A taxa de entidades ausentes é um indicador crítico que reflete a capacidade do modelo de identificar corretamente todas as entidades relevantes em um texto. A presença de uma alta taxa de M sugere que os modelos não estão conseguindo capturar todas as instâncias de entidades nomeadas, o que pode ser atribuído tanto a limitações nos dados de treinamento quanto à complexidade das entidades presentes nos textos literários.

Espúrias (S). As entidades espúrias referem-se a casos em que o modelo identifica incorretamente entidades que não estão presentes no texto. Esse tipo de erro pode ocorrer devido a ambiguidades linguísticas ou a uma modelagem inadequada de contextos. O número elevado de S, especialmente em cenários onde a precisão é crítica, indica a necessidade de ajustes finos e técnicas de pré-processamento mais robustas.

Incorretas (I). As entidades incorretas são aquelas que, embora o modelo tenha identificado uma entidade, a classificação atribuída está errada. Essa questão se torna evidente em classes de entidades mais complexas, como ORG e TIME, onde as expressões podem variar amplamente. Modelos que não conseguem lidar bem com a diversidade de referências literárias frequentemente apresentam taxas mais altas de I.

Parciais (P). As entidades parciais são aquelas em que apenas uma parte da entidade é identificada corretamente. Esse tipo de erro é especialmente comum em cenários com limites de entidades exigentes, onde a falta de precisão na segmentação pode levar a resultados insatisfatórios. A análise dos casos de P pode fornecer insights valiosos sobre como melhorar a segmentação e a identificação de entidades nos modelos.

Modelo	Real	Predito
LitBERTimbau	*Iam ser da Igreja, como seu dote.*	*Iam ser da Igreja, como seu dote.*
LitBERTimbau	*Nesta época eu era contratado pela Secretaria de Educação de Mato Grosso para dar aulas de teatro naquele estado, e resolvi utilizar meus alunos em laboratórios teatrais que tinham como tema a Táboa da Esmeralda.*	*Nesta época eu era contratado pela Secretaria de Educação de Mato Grosso para dar aulas de teatro naquele estado, e resolvi utilizar meus alunos em laboratórios teatrais que tinham como tema a Táboa da Esmeralda.*
BERT-CRF	*Era a hora que toda a Espanha dormia no verão.*	*Era a hora que toda a Espanha dormia no verão.*
BERT-CRF	*Sinhá Vitória acomodou os filhos, que arriaram como trouxas, cobriou-os com molambos.*	*Sinhá Vitória acomodou os filhos, que arriaram como trouxas, cobriou-os com molambos.*
BERT-CRF	*Nossa Senhora, com o Menino Jesus em seus braços, resolveu descer à Terra e visitar um mosteiro.*	*Nossa Senhora, com o Menino Jesus em seus braços, resolveu descer à Terra e visitar um mosteiro.*

ORG — LOC — PER — TIME

Tabela 10: Exemplos de entidades não detectadas por diferentes modelos.

7.1. Entidades Ausentes

Conforme a Figura 4, o modelo BERT-CRF apresenta a maior quantidade de entidades ausentes (312), indicando dificuldades significativas na captura de entidades nomeadas. Já os modelos ajustados mostram uma redução significativa nesse tipo de erro. O número elevado de entidades ausentes no modelo BERT-CRF puro pode ser atribuído à falta de adaptação ao domínio específico dos textos literários, o que sugere que a inclusão de dados de treinamento mais relevantes e específicos melhora substancialmente o desempenho dos modelos.

Como mostrado na Figura 3, de acordo com a métrica de revocação, os modelos ajustados FT BERT-CRF e LitBERT-CRF conseguiram capturar todas as poucas entidades ORG e TIME existentes no conjunto de teste. No entanto, os modelos BERT-CRF puro e LitBERTimbau apresentaram dificuldades em atingir o mesmo nível de desempenho. Já em relação às entidades PER e LOC, os três modelos ajustados apresentam um desempenho significativamente melhor em comparação com o *baseline*. A Tabela 10 mostra cinco exemplos que ilustram esses desafios.

Nos dois primeiros exemplos, o modelo LitBERTimbau falha ao capturar as entidades ORG *Igreja* e *Secretaria de Educação de Mato Grosso*. No caso da entidade *Igreja*, o LitBERTimbau não consegue reconhecer a entidade como uma organização e sim como pessoa, refletindo uma confusão entre os tipos de entidades que pode ser atribuída à complexidade do contexto literário. Essa falha na classificação correta das entidades pode ser devida à falta de exemplos suficientes de contextos semelhantes nos dados de treinamento do modelo, bem como ao desbalanceamento nas classes de entidades durante o treinamento, onde certas categorias podem ter sido super-representadas em relação a outras.

No segundo exemplo, o modelo LitBERTimbau reconhece parcialmente a entidade *Secretaria de Educação de Mato Grosso*, mas segmenta incorretamente *Mato*, identificando apenas *Secretaria de Educação* como uma entidade. Isso indica uma limitação na compreensão contextual do modelo, que pode ser melhorada com dados de treinamento mais específicos e detalhados sobre entidades compostas em contextos literários.

Já no terceiro exemplo, o modelo BERT-CRF puro não reconhece a entidade TIME *verão*, sugerindo que o modelo genérico não possui a capacidade suficiente para capturar certas expressões temporais comuns em textos literários. Esses exemplos destacam como a adaptação ao domínio e o ajuste fino podem melhorar significativamente a capacidade dos modelos de capturar entidades nomeadas em textos literários.

Os dois últimos exemplos mostram que o modelo BERT-CRF falha ao capturar as entidades PER *os filhos* e *o Menino Jesus*, bem como a entidade LOC *um mosteiro*. Essas falhas indicam que o modelo genérico enfrenta desafios significativos na identificação de entidades nomeadas que podem ser referenciadas de forma implícita ou menos convencional em textos literários. Por exemplo, a não detecção da entidade *os filhos* destaca uma possível limitação no reconhecimento de expressões que não possuem um nome próprio ou que são referidas de maneira mais genérica.

7.2. Entidades Espúrias

Já em relação às entidades espúrias, ou seja, aquelas que foram incorretamente identificadas como entidades pelo modelo, a análise revela particularidades importantes. No geral, ao comparar com o *baseline*, os modelos ajustados demonstraram um número considerável de entidades espúrias, sugerindo uma tendência a captar termos que não deveriam ser classificados como entidades nomeadas.

A Tabela 11 apresenta exemplos de entidades espúrias detectadas por diferentes modelos. O primeiro exemplo ilustra como o modelo pode errar ao identificar termos que, apesar de contextualizados, não têm relevância como entidades nomeadas. Por exemplo, os modelos FT BERT-CRF e LitBERTimbau identificaram a expressão *a bolandeira* como uma entidade, embora essa não represente uma organização, localização ou pessoa específica.

No entanto, os dois últimos exemplos destacam erros de anotação. No primeiro caso, os modelos LitBERT-CRF e LitBERTimbau rotularam corretamente a palavra *amigo* como entidade pessoa, apesar de, dependendo do contexto, o termo poder ser entendido como um substantivo comum. Já no segundo caso, os modelos FT BERT-CRF e LitBERTimbau classificam corretamente o termo *juazeiros* como uma entidade LOC. Contudo, como esse termo pode se referir a algo mais genérico, como a região ou um ponto de referência, é possível que sua classificação como LOC seja interpretada de forma ampla, mas precisa quando no contexto correto.

Esse tipo de confusão demonstra como o contexto e a polissemia das palavras podem impactar a precisão do reconhecimento de entidades nomeadas. A ambiguidade no uso da linguagem em textos literários é um desafio significativo para modelos de aprendizado de máquina, que dependem muitas vezes de padrões claros e bem definidos. Portanto, a inclusão de mais dados de treinamento que representem a complexidade e a diversidade da linguagem literária pode ajudar a reduzir esses erros, melhorando a precisão do modelo e sua capacidade de discernir entre diferentes usos das palavras.

7.3. Entidades Incorretas e Parciais

Para finalizar a análise de erros, também são discutidos casos de entidades incorretas, considerando os cenários **estrito**, **tipo** e **parcial**. As entidades incorretas referem-se a casos em que o modelo não apenas falha em detectar uma entidade, mas também classifica erroneamente um termo como uma entidade nomeada. Essa categoria de erro é crucial para entender as limitações dos modelos e suas capacidades de generalização.

No cenário **estrito**, onde a precisão e a recuperação de entidades devem ser rigorosamente observadas, a taxa de entidades incorretas tende a ser mais alta. Isso ocorre porque, nesse contexto, as entidades devem corresponder exatamente às referências verdadeiras no texto, e qualquer pequena divergência resulta em erro. Por exemplo, um modelo pode identificar corretamente a entidade *Universidade Federal* como uma organização, mas ao rotular *Universidade* como uma entidade isolada, ele falha em reconhecer a totalidade necessária da entidade, levando a uma classificação incorreta.

No cenário **tipo**, onde uma certa flexibilidade é permitida na sobreposição de entidades, os modelos ainda podem apresentar confusões entre categorias de entidades. Isso é evidente quando, por exemplo, uma entidade como *São Paulo* é classificada incorretamente como uma organização em vez de uma localização, devido à ambiguidade que pode surgir em textos literários que mencionam cidades em contextos específicos. Esses casos ressaltam a necessidade de um treinamento mais robusto e contextualizado, com dados que representem as variações de uso das entidades em diferentes contextos.

Por último, no cenário **parcial**, onde a correspondência dos limites é mais flexível, independentemente do tipo de entidade, os modelos tendem a apresentar uma taxa de detecção mais alta, mas ainda podem classificar entidades incorretamente. Nesse cenário, um modelo pode reconhecer um trecho do texto que contém parte de uma entidade, mas falhar ao identificar a entidade completa ou associá-la corretamente a sua categoria. Por exemplo, ao lidar com a expressão *Instituto Federal de Educação*, um modelo pode identificar apenas *Instituto* como

Modelo	Real	Predito
FT BERT-CRF LitBERTimbau	*Seu Tomás fugira também, com a seca, a bolandeira estava parada.*	*Seu Tomás fugira também, com a seca, a bolandeira estava parada.*
LitBERT-CRF LitBERTimbau	*Baleia jantara os pés, a cabeça, os ossos do amigo, e não guardava lembrança disto.*	*Baleia jantara os pés, a cabeça, os ossos do amigo, e não guardava lembrança disto.*
FT BERT-CRF LitBERTimbau	*Mas chegando aos juazeiros, encontrou os meninos adormecidos e não quis acordá-los.*	*Mas chegando aos juazeiros, encontrou os meninos adormecidos e não quis acordá-los.*

ORG — LOC — PER — TIME

Tabela 11: Exemplos de entidades espúrias detectadas por diferentes modelos.

Modelo	Real	Predito
FT BERT-CRF LitBERT-CRF	*Coitado, morrera na areia do rio [...]*	*Coitado, morrera na areia do rio [...]*
FT BERT-CRF LitBERT-CRF LitBERTimbau	*A cachorra Baleia foi enroscar-se junto dele.*	*A cachorra Baleia foi enroscar-se junto dele.*
FT BERT-CRF LitBERT-CRF LitBERTimbau	*Estavam no pátio de uma fazenda sem vida.*	*Estavam no pátio de uma fazenda sem vida.*

ORG — LOC — PER — TIME

Tabela 12: Exemplos de entidades incorretas detectadas por diferentes modelos.

uma entidade, desconsiderando o restante da expressão que completa a identificação.

A Tabela 12 fornece exemplos de entidades reconhecidas parcialmente, mostrando como os modelos lidam com casos em que a identificação é imprecisa ou incompleta. Os três exemplos ilustram que a maioria dos problemas de reconhecimento está relacionada à dificuldade dos modelos em captar a totalidade da entidade no contexto apresentado. Em todos os casos, os modelos demonstram dificuldades em capturar artigos e preposições, resultando em uma identificação fragmentada das entidades.

Esses casos destacam a importância de treinar os modelos com dados que incluam exemplos ricos e variados de estruturas linguísticas. Para melhorar o desempenho em reconhecimento de entidades, é fundamental que os modelos sejam capazes de lidar com a complexidade das construções linguísticas e a fluidez do idioma, garantindo assim uma representação mais fiel e abrangente das entidades presentes nos textos.

7.4. Discussão

No geral, os resultados mostraram que o modelo BERT-CRF puro apresentou a maior quantidade de entidades ausentes, evidenciando dificuldades significativas em capturar entidades nomeadas. Isso sugere que a falta de adaptação ao domínio específico dos textos literários é um fator crítico, que pode ser mitigado através da inclusão de dados de treinamento mais relevantes e específicos.

Já os modelos ajustados, como FT BERT-CRF, LitBERT-CRF e LitBERTimbau, apresentaram uma redução significativa nas entidades ausentes, demonstrando a eficácia do ajuste fino para melhorar o desempenho do modelo. No entanto, embora os modelos ajustados tenham uma tendência menor de identificar incorretamente termos como entidades, ainda há casos em que palavras contextualmente ambíguas são erradamente classificadas. Isso indica a necessidade de um treinamento mais rigoroso, que considere a complexidade dos textos literários.

8. Conclusão

Neste artigo, são investigadas estratégias de pré-treinamento adaptativas ao domínio para aprimorar a tarefa de reconhecimento de entidades nomeadas em textos literários em português. Para isso, dois modelos adaptativos ao domínio foram desenvolvidos, LitBERT-CRF e LitBERT-Timbau, construídos sobre modelos de linguagem de domínio geral, aproveitando dados literários para ajustar seus desempenhos.

Além de introduzir esses modelos, foi realizada uma análise comparativa, avaliando o aprendizado por transferência entre domínios em conjunto com um *baseline* de domínio geral. No geral, ambos os modelos adaptativos ao domínio superam o modelo *baseline* BERT-CRF, mostrando os potenciais benefícios de incorporar dados específicos do domínio no processo de pré-treinamento. Em particular, o LitBERT-CRF supera os outros modelos avaliados, apresentando resultados competitivos em diferentes cenários de avaliação, destacando-se na identificação estrita de entidades literárias.

Nosso estudo também destacou os *trade-offs* associados a diferentes estratégias adaptativas ao domínio. O modelo de aprendizado por transferência entre domínios (FT BERT-CRF) mostrou resultados competitivos, especialmente em cenários de avaliação onde a correspondência exata de limites não é necessária. Os modelos de pré-treinamento adaptativo ao domínio, que incorporam dados literários no processo de pré-treinamento, mostraram uma precisão superior no reconhecimento de entidades literárias.

Por fim, a análise de erros revelou que modelos ajustados, como FT BERT-CRF e LitBERT-CRF, apresentaram uma redução significativa nas entidades ausentes em comparação com o modelo BERT-CRF puro, sugerindo que a adaptação ao domínio é crucial para melhorar o desempenho do modelo. No entanto, ainda foram observadas dificuldades em capturar a totalidade das entidades, especialmente em estruturas complexas. A análise também destacou a necessidade de um treinamento mais equilibrado para reduzir a identificação de entidades espúrias, que foram mais prevalentes nos modelos ajustados.

Limitações. Apesar dos avanços, o presente estudo apresenta algumas limitações. A principal limitação reside na disponibilidade e diversidade dos dados literários utilizados no pré-treinamento. O desbalanceamento substancial entre as classes de entidades nomeadas durante o treinamento pode ter influenciado o desempenho dos modelos, particularmente em categorias menos representadas, como ORG e TIME. Além disso, a análise de erros indicou que os modelos ainda têm dificuldades em capturar artigos e preposições, o que pode ser atribuído tanto à complexidade inerente do texto literário quanto ao processo de anotação manual, que pode introduzir inconsistências na identificação das entidades.

Outra limitação importante diz respeito à diferença na conceituação de entidades nomeadas nos corpora *PPORTAL_ner* e HAREM. As distinções nas definições e no detalhamento das entidades entre esses conjuntos podem ter afetado a capacidade dos modelos de generalizar adequadamente para o contexto literário, uma vez que as categorias e anotações podem não ter sido totalmente compatíveis com as especificidades dos textos analisados. Por fim, há uma concentração de obras no século XIX e início do século XX nos corpora utilizados, o que pode introduzir distorções linguísticas, particularmente no vocabulário e nas atualizações linguísticas, limitando a generalização dos modelos para textos mais modernos ou de outros períodos.

Trabalhos Futuros. Nossas descobertas abrem várias vias para investigações futuras. Por exemplo, incorporar um conjunto mais extenso e diversificado de corpora literários pode ajudar a capturar uma gama mais ampla de nuances linguísticas e melhorar a robustez dos modelos. Pesquisas futuras também podem investigar a otimização de hiperparâmetros e protocolos de treinamento avançados para ajustar os modelos de forma mais eficaz. Além disso, explorar outras tarefas dentro do domínio literário, como análise de sentimento, classificação de texto ou tarefas multilíngues, pode fornecer insights sobre a versatilidade e robustez dos modelos avaliados.

Outro aspecto relevante para futuros trabalhos é o alinhamento dos modelos com a definição de entidades nomeadas do conjunto *PPORTAL_ner*. Isso poderia melhorar a consistência e a generalização dos modelos, permitindo que eles lidem melhor com as especificidades dos textos literários e ajustem suas abordagens para as particularidades do domínio. Além disso, pretendemos realizar uma comparação de técnicas de adaptação mais avançadas, como adaptação de domínios e de tarefas para transformers, que podem oferecer melhores resultados em termos de eficiência computacional e desempenho. Uma linha de pesquisa adicional seria expandir a diversidade temporal dos corpora utilizados no pré-treinamento e ajuste fino, incluindo obras de períodos mais recentes e de diferentes gêneros literários. Essa abordagem ajudaria a mitigar o viés temporal presente nos corpora atuais.

Agradecimentos

Este trabalho foi parcialmente financiado pela CAPES, CNPq e FAPEMIG, Brasil.

Disponibilidade de Dados

Os conjuntos de dados gerados e/ou analisados durante o presente estudo estão disponíveis online (Silva & Moro, 2024c). Os models de linguagem desenvolvidos neste trabalho estão disponíveis em https://huggingface.co/marianaossilva.

Referências

Bamman, David, Sejal Popat & Sheng Shen. 2019. An annotated dataset of literary entities. Em *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2138–2144. 10.18653/v1/n19-1220

Beltagy, Iz, Kyle Lo & Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. Em *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*, 3613–3618. 10.18653/V1/D19-1371

Bick, Eckhard. 2000. *The parsing system palavras: Automatic grammatical analysis of portuguese in a constraint grammar famework*. Aarhus University Press

Boukkouri, Hicham El, Olivier Ferret, Thomas Lavergne & Pierre Zweigenbaum. 2022. Retrain or train from scratch? comparing pretraining strategies of BERT in the medical domain. Em *13th Language Resources and Evaluation Conference (LREC)*, 2626–2633.

Claro, Daniela Barreiro, Joaquim Santos, Marlo Souza, Renata Vieira & Vládia Pinheiro. 2023. Extração de informação. Em *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, chap. 17. BPLN.

Cui, Shengmin & Inwhee Joe. 2023. A multi-head adjacent attention-based pyramid layered model for nested named entity recognition. *Neural Computing and Applications* 35(3). 2561–2574. 10.1007/S00521-022-07747-8

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. Em *Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186. 10.18653/V1/N19-1423

Docío, Susana Sotelo, Pablo Gamallo & Álvaro Iriarte. 2023. Desenvolvimento e avaliação de um modelo NER no domínio da análise cultural e do turismo. *Linguamática* 15(2). 3–18. 10.21814/lm.15.2.405

Doddington, George R., Alexis Mitchell, Mark A. Przybocki, Lance A. Ramshaw, Stephanie M. Strassel & Ralph M. Weischedel. 2004. The automatic content extraction (ACE) program - tasks, data, and evaluation. Em *4th International Conference on Language Resources and Evaluation (LREC)*, 837–840.

Emelyanov, Anton A. & Ekaterina Artemova. 2019. Multilingual named entity recognition using pretrained embeddings, attention mechanism and NCRF. Em *7th Workshop on Balto-Slavic Natural Language Processing*, 94–99. 10.18653/V1/W19-3713

Filho, Jorge A. Wagner, Rodrigo Wilkens, Marco Idiart & Aline Villavicencio. 2018. The brWaC corpus: A new open resource for Brazilian Portuguese. Em *International Conference on Language Resources and Evaluation*, 4339–4344.

Freitas, Cláudia, Elvis Souza, Maria Clara Castro, Tatiana Cavalcanti, Patricia Ferreira da Silva & Fábio Corrêa Cordeiro. 2023. Recursos linguísticos para o PLN específico de domínio: o Petrolês. *Linguamática* 15(2). 51–68. 10.21814/lm.15.2.412

Frontini, Francesca, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos & Ranka Stanković. 2020. Named entity recognition for distant reading in ELTeC. Em *CLARIN Annual Conference*, 37–41.

Gururangan, Suchin, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey & Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. Em *Annual Meeting of the Association for Computational Linguistics*, 8342–8360. 10.18653/V1/2020.ACL-MAIN.740

Lamproudis, Anastasios & Aron Henriksson. 2022. On the impact of the vocabulary for domain-adaptive pretraining of clinical language models. Em *15th International Joint COnference on Biomedical Engineering Systems and Technologies*, 315–332. 10.1007/978-3-031-38854-5_16

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So & Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36(4). 1234–1240. 10.1093/bioinformatics/btz682

Li, Jing, Aixin Sun, Jianglei Han & Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering* 34(1). 50–70. 10.1109/TKDE.2020.2981314

Liu, Zhuang, Degen Huang, Kaiyu Huang, Zhuang Li & Jun Zhao. 2020. FinBERT: A pre-trained financial language representation model for financial text mining. Em *29th International Joint Conference on Artificial Intelligence*, 4513–4519. 10.24963/IJCAI.2020/622

Mou, Lili, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang & Zhi Jin. 2016. How transferable are neural networks in NLP applications? Em *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 479–489. 10.18653/v1/D16-1046

de Oliveira, Lucas Ferro Antunes, Adriana S. Pagano, Lucas Emanuel Silva e Oliveira & Claudia Moro. 2022. Challenges in annotating a treebank of clinical narratives in Brazilian Portuguese. Em *Conference on Computational Processing of the Portuguese Language (PROPOR)*, 90–100. 10.1007/978-3-030-98305-5_9

Qiu, XiPeng, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai & XuanJing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences* 63(10). 1872–1897. 10.1007/s11431-020-1647-3

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li & Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)* 21. 140:1–140:67.

Rodrigues, Rafael BM, Pedro Ivo Monteiro Privatto, Gustavo José de Sousa, Rafael P. Murari, Luis C. S. Afonso, João P. Papa, Daniel C. G. Pedronette, Ivan Rizzo Guilherme, Stephan R. Perrout & Aliel F. Riente. 2022. PetroBERT: A domain adaptation language model for oil and gas applications in Portuguese. Em *Conference on Computational Processing of the Portuguese Language (PROPOR)*, 101–109. 10.1007/978-3-030-98305-5_10

Rodríguez, Dalia Andrea, Julia Diaz-Escobar, Arnoldo Díaz-Ramírez & Leonardo Trujillo. 2023. Domain-adaptive pre-training on a BERT model for the automatic detection of misogynistic tweets in Spanish. *Social Network Analysis and Mining* 13. 126. 10.1007/S13278-023-01128-2

Santos, Diana, Eckhard Bick & Marcin Wlodek. 2020. Avaliando entidades mencionadas na coleção ELTeC-por. *Linguamática* 12(2). 29–49. 10.21814/lm.12.2.336

Santos, Diana, Nuno Seco, Nuno Cardoso & Rui Vilela. 2006. HAREM: An advanced NER evaluation contest for Portuguese. Em *Conference on Language Resources and Evaluation (LREC)*, 1986–1991.

Santos, Diana, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, Cristina Mota, Emanoel Pires, Rebeca Schumacher Fuão & Paulo Silva Pereira. 2022. Identifying literary characters in Portuguese - challenges of an international shared task. Em *Conference on Computational Processing of the Portuguese Language (PROPOR)*, 413–419. 10.1007/978-3-030-98305-5_39

Segura-Bedmar, Isabel, Paloma Martínez & María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). Em *International Workshop on Semantic Evaluation (SemEval)*, 341–350.

Silva, Mariana O. & Mirella M. Moro. 2024a. Evaluating pre-training strategies for literary named entity recognition in Portuguese. Em *Conference on Computational Processing of Portuguese (PROPOR)*, 384–393.

Silva, Mariana O. & Mirella M. Moro. 2024b. PPORTAL_ner: An annotated corpus of Portuguese literary entities. Em *Procs Joint Int'l Conf on Computational Linguistics, Language Resources and Evaluation*, 12927–12937. ELRA and ICCL.

Silva, Mariana O. & Mirella M. Moro. 2024c. PPORTAL_ner: an annotated corpus of Portuguese literary entities. Zenodo. 10.5281/zenodo.10855187

Silva, Mariana O., Clarisse Scofield, Luiza de Melo-Gomes & Mirella M. Moro. 2022. Cross-collection dataset of public domain Portuguese-language works. *Journal of Information and Data Management* 13(1). 10.5753/jidm.2022.2349

Silva, Mariana O., Clarisse Scofield & Mirella M. Moro. 2021. PPORTAL: Public domain Portuguese-language literature dataset. Em *III Dataset Showcase Workshop*, 77–88. 10.5753/dsw.2021.17416

Singhal, Peeyush, Rahee Walambe, Sheela Ramanna & Ketan Kotecha. 2023. Domain adaptation: Challenges, methods, datasets, and applications. *IEEE Access* 11. 6973–7020. 10.1109/ACCESS.2023.3237025

Souza, Fábio, Rodrigo Frassetto Nogueira & Roberto de Alencar Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. ArXiv [cs.CL]. 10.48550/arXiv.1909.10649

Souza, Fábio, Rodrigo Frassetto Nogueira & Roberto de Alencar Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. Em *9th Brazilian Conference on Intelligent Systems*, 403–417. 10.1007/978-3-030-61377-8_28

Vieira, Andressa & Silva. 2023. Uma revisão para o reconhecimento de entidades nomeadas aplicado à língua portuguesa. *Linguamática* 15(2). 69–85. 10.21814/lm.15.2.396

Zhuang, Fuzhen, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong & Qing He. 2021. A comprehensive survey on transfer learning. *Proceedings of the IEEE* 109(1). 43–76. 10.1109/JPROC.2020.3004555