PPORTAL_ner: An Annotated Corpus of Portuguese Literary Entities

Mariana O. Silva, Mirella M. Moro
Computer Science Department, Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
{mariana.santos, mirella}@dcc.ufmg.br

Abstract

The intersection of natural language processing (NLP) and literary analysis has yielded valuable insights and applications across various languages. However, the scarcity of labeled data tailored for Portuguese literary texts poses a notable challenge. To address this gap, we present the *PPORTAL_ner* corpus, an annotated dataset that simplifies the development of Named Entity Recognition (NER) models specifically adapted for Portuguese literary works. Our corpus includes annotations of PER, LOC, GPE, ORG, and DATE entities within a diverse set of 25 literary texts. Annotation of the corpus involved a two-step process: initial pre-annotation using a pre-trained spaCy model followed by correction and refinement using the Prodigy annotation tool. With a total of 125,059 tokens and 5,266 annotated entities, *PPORTAL_ner* corpus significantly enriches the landscape of resources available for computational literary analysis in Portuguese. This paper details the annotation methodology, guidelines, and dataset statistics while also evaluating four NER models over the *PPORTAL_ner* corpus. Our evaluation analysis reveals that fine-tuning on domain-specific data significantly improves NER model performance, demonstrating the value of the *PPORTAL_ner* corpus for developing domain-specific language models.

Keywords: named entity recognition, corpus annotation, Portuguese literature, literary texts

1. Introduction

Fundamental questions concerning characters (*who?*), settings (*where?*), and temporal dimensions (*when?*) within literary works need the accurate extraction of entities, laying the foundation for insightful literary analysis. However, a challenge in the domain of computational literary analysis is the lack of labeled data for training Named Entity Recognition (NER) models specifically tailored to literary texts. As a result, most analysis endeavors rely on NER models trained on generic corpora, such as news articles, which do not align with the unique characteristics of literary texts.

Unlike news texts, which are oriented toward geopolitical entities and organizations, literary works offer quite different distributions of entity categories (Bamman et al., 2019). They place a strong emphasis on individuals, such as characters, and provide rich and detailed descriptions of various settings. Such diversity in entity types sets literary texts apart, rendering them a distinctive and intriguing domain for exploration, then requiring the development of specialized NER models and datasets for the literary domain.

The need for labeled data and tailored NER models for literary texts entails further challenges and complications in computational literary analysis. First, models trained on out-of-domain data are likely to experience significant performance degradation when applied to literary texts, primarily due to the vast differences in entity distribution (Lamproudis et al., 2022). Second, without in-domain test data, it becomes challenging to quantify the extent of this performance degradation precisely (Singhal et al., 2023). Consequently, the need for high-quality, domain-specific annotated datasets and specialized NER models becomes evident.

Furthermore, when delving into the Portuguese-language literature domain, the research gap becomes even more pronounced (Grilo et al., 2020; Albuquerque et al., 2023). While the challenges of entity recognition in literary texts are well-documented in English, they add complexity when dealing with a language as rich and nuanced as Portuguese. The peculiarities of the Portuguese language, including its intricate grammar and subtle variations in meaning, demand specialized attention. Such intricacies pose distinct challenges that cannot adequately be addressed by using models trained on other languages or general-domain datasets.

Our effort to bridge this research gap and contribute to the field of computational literary analysis focuses on Portuguese literary texts. By creating an annotated dataset of Portuguese literary entities, called *PPORTAL_ner*, we aim to provide researchers and NER model developers with a valuable resource for the intricacies of the Portuguese language and its rich literary tradition.

PPORTAL_ner corpus is tailored to Brazilian and Portuguese literary texts, containing annotations for five entity categories, including PER, LOC, GPE, ORG, and DATE. Within a diverse collection of 25 literary works, it contains 125,059 tokens and 5,266 annotated entities. This dataset contributes to the development of potentially more accurate and

LREC-COLING 2024, pages 12927–12937

Table 1: A summary of existing corpora in the literary domain.

Reference	Corpus	Language	Category	Size*
(Bamman et al., 2019)	LitBank	English	Literary entities	100 (2,000 tokens each)
(Dekker et al., 2019)	OWTO	English	Literary entities	40 (300 sentences each)
(Bamman et al., 2020)	LitBank	English	Literary coreference	100 (2,000 tokens each)
(Papay and Padó, 2020)	RiQuA	English	Literary quotation	11 (full-length)
(Grilo et al., 2020)	BDCamões	Portuguese	Literary corpus	200 (full-length)
(Stymne and Östman, 2020)	SLäNDa	Swedish	Literary quotation	8 (2–10 chapters each)
(Santos, 2021)	ELTeC-por	Portuguese	Literary entities	100 (full-length)
(Vishnubhotla et al., 2022)	PDNC	English	Literary quotation	22 (full-length)
(Stymne and Östman, 2022)	SLäNDa 2.0	Swedish	Literary quotation	19 (full-length)
(Morgado et al., 2022)	MEGALITE-PT	Portuguese	Literary corpus	4,311 (full-length)
(Silva and Moro, 2024b)	***PPORTAL_ner***	Portuguese**	Literary entities	25 (5,000 tokens each)

*Total of literary works considered
**Brazil's and Portugal's

context-aware NER models and encourages further exploration within Portuguese literature. The dataset is freely available for download at (Silva and Moro, 2024b).

2. Related Work

Despite the significant growth in computational literary research (Frontini et al., 2020), the availability of labeled data suitable for training NER models tailored for literary texts remains deficient (Silva and Moro, 2024a). Such shortage of data has limited the development and evaluation of NER models in the context of literature (Bamman et al., 2019), increasing the complexity of addressing this domain's unique requirements and nuances.

In such a context, Bamman et al. (2019) introduced LitBank, a dataset comprising entity annotations from 100 public-domain works sourced from Project Gutenberg[1] and written in English. LitBank covers six of the ACE 2005 (LDC, 2005) entity categories: people (PER), facilities (FAC), geo-political entities (GPE), locations (LOC), vehicles (VEH), and organizations (ORG). Their findings demonstrate the effectiveness of training nested entity recognition models on in-domain literary data, resulting in a substantial improvement in F-score.

Similarly, Dekker et al. (2019) created the OWTO dataset, a valuable resource with 40 English novels–20 new and 20 old. OWTO focuses on the entity type "person". While there are other English annotated datasets in the literary domain, they often target different aspects of text analysis, such as quotation attribution (Papay and Padó, 2020; Stymne and Östman, 2022; Vishnubhotla et al., 2022) and coreference (Bamman et al., 2020). These resources are invaluable for various NLP tasks but may not directly address the dearth of labeled data for NER in the context of literary texts.

The data gap becomes even more pronounced in non-English literature, such as the Portuguese one (which includes both Brazilian and Portuguese authors). Despite the existence of corpora for Portuguese-language literary works (Grilo et al., 2020; Morgado et al., 2022), the shortage of *annotated literary entities*' data remains a pervasive issue (Frontini et al., 2020; Schöch et al., 2021). Such a lack poses challenges for developing NER models tailored to the distinct characteristics of literary works written in Portuguese.

Table 1 provides an overview of the aforementioned corpora in the literary domain, highlighting key characteristics such as language, category, and size. While most of these resources are centered around English literary materials, they explore various facets of literary analysis, including entities, coreference, and quotations. Notably, only three corpora are dedicated to Portuguese literary texts, with two of them, BDCamões (Grilo et al., 2020) and MEGALITE-PT (Morgado et al., 2022), offering raw text without named entity annotations.

The ELTeC-por corpus, part of the European Literary Text Collection (ELTeC),[2] stands as the only corpus of literary entity annotations in Portuguese. Focused on public-domain novels released between 1840 and 1920, it contains 100 literary works in European Portuguese. Using the PALAVRAS-NER system (Bick, 2006), the ELTeC-por corpus contains literary entities annotation spanning six classes, including person names (PERS), organizations (ORG), locations (LOC), events (EVENT), works of art (WORK), and brands (BRAND).

In contrast, our *PPORTAL_ner* corpus offers a comprehensive collection of literary works considering both Brazilian and Portuguese literature, thereby enhancing the diversity of the corpus. However, it is important to note that our corpus may present a more limited size compared to ELTeC-

[1] https://gutenberg.org/

[2] https://distant-reading.net

por, yet it compensates by covering a wider temporal and geographical scope within the Portuguese literary landscape.

3. Data

Our corpus is sourced from *PPORTAL*,[3] an extensive repository of metadata containing over 80,000 public domain literary works in the Portuguese language, predominantly derived from Brazil and Portugal (Silva et al., 2021, 2022). *PPORTAL* aggregates data from three digital libraries: Domínio Público,[4] Projecto Adamastor,[5] and Biblioteca Digital de Literatura dos Países Lusófonos (BLPL).[6]

To simplify referencing, this new dataset is called *PPORTAL_ner*. Its selection process contains a diverse range of 25 individual literary works, spanning different authors and literary styles. All of these texts were published prior to 1953, adhering to the current criteria for public domain status in Brazil, with the majority falling within the timeframe spanning from 1554 to 1938.

To ensure the uniformity and quality of the corpus, each text underwent pre-processing, removing special characters (excluding hyphens and punctuation marks, given their relevance in literary contexts). Additionally, this pre-processing step removed emails and website references. Next, we extracted a representative sample from each text, amounting to approximately the initial 5,000 words, equivalent to two chapters on average. Such a curation process resulted in a comprehensive dataset of 125,014 tokens.

4. Annotation

Our annotation guidelines drew inspiration from the fundamental literary questions concerning the characters (*who?*), settings (*where?*), and temporal dimensions (*when?*). These questions served as the guiding principles for our annotation process, as they support the central elements of literary analysis in Portuguese literary texts.

4.1. Entity Categories

The annotation process focuses on five essential entity categories commonly employed in most NER systems. While we recognize that the literary context may demand additional categories to capture the complexity and richness of entities within these texts, we chose five categories that align with these fundamental questions and allow us to evaluate NER models considering traditional NE categories.

PER. Dedicated to labeling entities representing individuals within the text. These individuals can be explicitly referred to by their proper names, such as “Capitu”. Nominal references are also included, which are anaphoric noun phrases referring to characters (e.g., “Bentinho’s wife”). Furthermore, this category also contains sets of people, such as family units (e.g., “Capitu’s parents”), acknowledging the diverse ways individuals are referenced and portrayed in literary texts.

LOC. Dedicated to labeling entities that denote locations, geographical landmarks, and references to specific places within the text. Note that this class distinguishes itself from the GPE category, which primarily pertains to real-world geographical locations (e.g., London, New York) and nations (e.g., England, the United States). In contrast, the LOC category contains both named and common imagined entities within the literary context, such as “the town” or “the village”.

GPE. Dedicated to labeling terms denoting the text’s geopolitical entities. These entities entail a broad range, including countries and cities.

ORG. Dedicated to labeling entities that denote organizations, institutions, and groups mentioned within the texts. In literary data, *organizations* are defined as formal associations and are relatively rare, being the least frequently occurring entity class. Among the instances of organizations in the dataset, the most recurrent examples include references to entities like the Church or the police.

DATE. Dedicated to annotations related to dates, time periods, or temporal expressions found within the literary works.

4.2. Annotation Guidelines

Our annotation process is executed by following a well-defined set of guidelines to ensure consistency and accuracy in entity labeling. These guidelines follow a two-step approach that involves initial pre-annotation using the pre-trained spaCy model *pt_core_news_lg* and subsequent correction and refinement using the Prodigy annotation tool.[7]

4.2.1. Pre-Annotation

Initially, all 25 literary texts are pre-annotated using the spaCy model `pt_core_news_lg`, a pre-trained language model specifically designed for Portuguese. The spaCy model provided initial entity suggestions based on its knowledge of the language and entity recognition capabilities.

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

[4] https://www.dominiopublico.gov.br

[5] https://projectoadamastor.org

[6] https://www.literaturabrasileira.ufsc.br

[7] https://prodi.gy/

4.2.2. Correction and Refinement

The Prodigy annotation tool is employed to correct and refine pre-annotated data. The `ner.correct` recipe in Prodigy facilitated the creation of a gold-standard dataset. Within this recipe, we leverage the `-update` argument to continuously update the model during the annotation loop, allowing it to adapt and improve based on received annotations. Furthermore, we maintain the boolean segmentation argument set to `true`, facilitating the segmentation of lengthy literary texts into sentences, thereby streamlining the annotation process.

The annotation guidelines emphasized the following principles to guide one author annotator through the correction process.

Accuracy. The annotator was instructed to prioritize accuracy when reviewing and revising entity labels, ensuring that entities were correctly identified and categorized.

Consistency. Consistency in labeling was crucial throughout the dataset. The annotator was encouraged to maintain consistent entity labeling conventions, following the guidelines established by the spaCy model where applicable.

Context Awareness. The annotator was advised to consider the broader context of the text when making corrections, especially for entities with complex or ambiguous references.

Nested Entities. The annotator was directed to recognize and label nested entities as a flat structure in which entity labels cannot be embedded within each other. For example, in the following sentence, the annotator should recognize that "Capitu's parents" is a hierarchical entity consisting of two individuals, Senhor Pádua and Dona Fortunata. To maintain a flat structure, each individual is annotated separately without embedding entity labels within each other:

PER PER

Capitu's parents, *Sr. Pádua and Dona Fortunata, were concerned about their children.*

Distinguishing GPE and LOC. Clear guidance was provided for distinguishing between Geopolitical Entities (GPE) and Locations (LOC), with a focus on real-world geographic entities for GPE and a broader scope for LOC, considering both named and common imagined places.

4.3. Annotation Format

The annotations are in JSON format, including the document ID (`doc_id`), document text (`doc_text`), and a list of annotated entities (`entities`). Each entity object in the list contains information such as entity ID (`entity_id`), entity text (`text`), entity label (`label`), start offset (`start_offset`), and end offset (`end_offset`). An example is presented as follows:

```
{
  "doc_id": 2550,
  "doc_text": "Se a lembrança de Iracema
  estivesse nalma do estrangeiro, ela
  não o deixaria partir." (Portuguese)
  "If the memory of Iracema was in the
  foreigner's mind, she wouldn't let
  him go.",
  "entities": [
    {
      "entity_id": 1,
      "text": "Iracema",
      "label": "PESSOA",
      "start_offset": 18,
      "end_offset": 25
    },
    {
      "entity_id": 2,
      "text": "estrangeiro",
      "label": "PESSOA",
      "start_offset": 45,
      "end_offset": 56
    }
  ]
}
```

The JSON format is not exclusive to spaCy, making it versatile for different NLP tools and platforms. Its common structure enhances the corpus's compatibility, facilitating seamless integration into different research workflows and applications.

5. *PPORTAL_ner* Statistics

In this section, we present key statistics and characteristics of the *PPORTAL_ner* corpus. Table 2 summarizes these statistics, including metadata and the number of tokens, sentences and annotated entities.

5.1. Metadata Information

In addition to the textual content, the *PPORTAL_ner* corpus provides metadata for each literary work, including the author's name, title, language, and publication year. As presented in Table 2, our corpus showcases a diverse array of literary works spanning several centuries and featuring authors from Portuguese and Brazilian literature. The metadata serves as a valuable resource for researchers, allowing them to explore the corpus with a historical and authorship perspective, further enriching the potential applications of our dataset.

5.2. Corpus Size

The corpus contains 25 individual literary works selected to offer a broad and diverse perspective on the landscape of Portuguese-written literature.

Table 2: Corpus main statistics.

Title	Author	Lang.	Year	#T	#S	#E
Menina e Moça	Bernardim Ribeiro	PT	1554	5,004	139	106
Os Lusíadas	Luís Vaz de Camões	PT	1572	5,000	118	188
Eurico, o Presbítero	Alexandre Herculano	PT	1844	5,000	147	174
Memórias de um Sargento de Milícias	Manuel Antônio de Almeida	PT-BR	1854	5,000	166	212
Amor de Perdição	Camilo Castelo Branco	PT	1861	5,001	204	309
Iracema	José de Alencar	PT-BR	1865	5,000	275	335
As Pupilas do Senhor Reitor	Júlio Dinis	PT	1867	5,000	264	199
A Morgadinha dos Canaviais	Júlio Dinis	PT	1868	5,000	245	242
Inocência	Visconde de Taunay	PT-BR	1872	5,002	204	156
Helena	Machado de Assis	PT-BR	1876	5,000	275	260
O Mandarim	Eça de Queirós	PT	1880	5,000	144	180
Memórias Póstumas de Brás Cubas	Machado de Assis	PT-BR	1881	5,002	204	156
O Alienista	Machado de Assis	PT-BR	1882	5,001	195	218
Casa de Pensão	Aluísio Azevedo	PT-BR	1884	5,000	289	254
O Cortiço	Aluísio Azevedo	PT-BR	1890	5,000	195	153
Quincas Borba	Machado de Assis	PT-BR	1891	5,000	299	201
Dom Casmurro	Machado de Assis	PT-BR	1899	5,003	256	183
Os Sertões	Euclides da Cunha	PT-BR	1902	5,000	154	179
Esaú e Jacó	Machado de Assis	PT-BR	1904	5,000	249	204
Cartas de Inglaterra	Eça de Queirós	PT	1905	5,000	134	299
Memorial de Aires	Machado de Assis	PT-BR	1908	5,001	271	245
A Confissão de Lúcio	Mário de Sá-Carneiro	PT	1913	5,000	244	137
Alves & Companhia	Eça de Queirós	PT	1925	5,000	192	172
Capitães da Areia	Jorge Amado	PT-BR	1937	5,001	225	306
Vidas Secas	Graciliano Ramos	PT-BR	1938	5,001	345	230
	Total			**125,014**	**5,418**	**5,266**

Lang.: Language | #T: Total of tokens | #S: Total of sentences | #E: Total of entities

Table 3: Distribution of entity categories.

Category	Frequency (%)	Examples
PER	3,609 (68.53%)	“Capitu”, “the foreigner”, “the youngest son”
LOC	1,126 (21.38%)	“the village”, “the town”, “under the bridge”
GPE	315 (5.98%)	“Brazil”, “Lisbon”, “Rio de Janeiro”, “Europe”
ORG	115 (2.18%)	“the police”, “the Church”, “the army”
DATE	101 (1.92%)	“XVIII century”, “1847”, “the winter”

In total, the corpus contains 125,059 tokens, 5,418 sentences, and 5,266 annotated entities.

5.3. Entity Distribution

The *PPORTAL_ner* corpus exhibits a diverse distribution of annotated entities across five distinct categories: PER, LOC, GPE, ORG, and DATE. Table 3 provides a comprehensive breakdown of each entity category’s frequency, expressed as a percentage of the total annotated entities, along with illustrative examples that offer a glimpse into the dataset’s content.

As previously discussed, one notable characteristic of literary texts is their distinctive distribution of entity categories, which deviates from more news-centric datasets. In our corpus, there is a pronounced emphasis on entities related to individuals (PER) and vivid descriptions of places (LOC). Both categories collectively account for nearly 90% of all annotated entities in the dataset. This unique emphasis aligns with the thematic focus of Portuguese literary works, where character portrayal and immersive settings play a central role in storytelling.

6. Evaluation

Adopting general-domain models in specialized domains often faces suboptimal performance due to significant domain-specific differences (Singhal et al., 2023). Therefore, researchers have explored strategies to develop domain-specific language models (Bamman et al., 2019), leveraging techniques such as continuous domain adaptive pre-training or fine-tuning on domain-specific corpora to adapt existing generic models to target domains (Singhal et al., 2023).

In this section, we fine-tuned four pre-trained

Table 4: Main characteristics of the considered pre-trained models.

Model	Pre-training Dataset	Domain	Entities Tags
pt_core_news_sm	WikiNER 2017	news, media	PER, LOC, ORG, MISC
pt_core_news_md	WikiNER 2017	news, media	PER, LOC, ORG, MISC
pt_core_news_lg	WikiNER 2017	news, media	PER, LOC, ORG, MISC
BERT-CRF	HAREM (Selective) 2006	general	PER, LOC, ORG, VALUE, TIME

models to assess our corpus's potential contributions to domain-specific language modeling and NLP tasks. We then compared their original versions to assess the extent of performance enhancement facilitated by our corpus.

6.1. Pre-trained Models

Table 4 shows the main characteristics of each considered pre-trained model. In particular, we consider three pre-trained models from spaCy's library. SpaCy offers pre-trained NER models for Portuguese in different sizes, including *large*, *medium*, and *small*. For this study, all three sizes were considered. Note that these spaCy models were originally trained on the WikiNER annotation (Ghaddar and Langlais, 2017), which does not contain all the entity classes in our corpus. To align the spaCy annotations with our gold standard, we made adjustments, considering GPE instances as LOC and DATE instances as MISC.

In addition to spaCy models, we assessed a BERT-based model introduced in (Souza et al., 2019). Souza et al. proposed the BERT-CRF model that combines a BERT-based embedding model with a Conditional Random Fields layer. Built upon the BERTimbau (Souza et al., 2020), a Portuguese-tailored BERT-based embedding model, BERT-CRF was initially trained on the HAREM corpus (Santos et al., 2006). The HAREM corpus offers two versions, and we utilized the "selective" version, featuring five classes: Person, Organization, Location, Value, and Time. We harmonized this version with our gold standard, treating GPE entities as LOC and DATE entities as TIME.

6.2. Experimental Setup

The outcome of the annotation process results in a collection of 6,965 annotated sentences originating from 25 distinct literary works. To facilitate the evaluation of the four language models, we employ a sentence-level stratification approach, partitioning the annotated sentences into training, development, and test sets. This stratification involves allocating 80% of the sentences to the training set, equivalent to 5,572 sentences. Furthermore, 10% of the sentences, amounting to 696 sentences, are designated for the validation set, while the remaining 10%, totaling 697 sentences, constitute the test set.

Table 5: NER models evaluation results on different training data.

Model	Training Data	P	R	F1
pt_core_news_sm	WikiNER	0.44	0.22	0.29
	PPORTAL_ner	0.67	0.49	0.56
pt_core_news_md	WikiNER	0.49	0.24	0.32
	PPORTAL_ner	0.66	0.52	0.58
pt_core_news_lg	WikiNER	0.47	0.23	0.31
	PPORTAL_ner	0.69	0.60	0.64
BERT-CRF	HAREM	0.79	0.27	0.41
	PPORTAL_ner	0.77	0.77	0.77

P: Precision | R: Recall | F1: F1 Score

All four pre-trained models are fine-tuned on the NER downstream task using our literary annotated corpus (*PPORTAL_ner*).[8] During fine-tuning, all models are trained for a fixed number of 10 epochs. No extensive hyperparameter search is performed, as the primary objective is to compare and evaluate the domain adaptation strategy for creating literary language models rather than achieving state-of-the-art performance on downstream tasks. We employ a narrow hyperparameter search with predefined parameters to evaluate the models fairly. In these fine-tuning sessions, the models are trained until they converge in terms of the validation set loss, ensuring that they reach a stable performance level.

6.3. Results and Discussion

The performance of the pre-trained NER models is presented in Table 5. We evaluated four different models, varying the training data to verify whether fine-tuning on domain-specific corpora (in this case, literary texts) significantly enhances the performance of the pre-trained models. The goal is to ascertain whether adapting to a domain could significantly enhance the models' capability to recognize named entities in literary texts.

[8]There is no comparison with the existing corpora for Portuguese literature because most of them provide only raw text, without annotations. Regarding the ELTeC-por corpus, a direct comparison with our results is left for future work due to differences in format and classes in annotations.

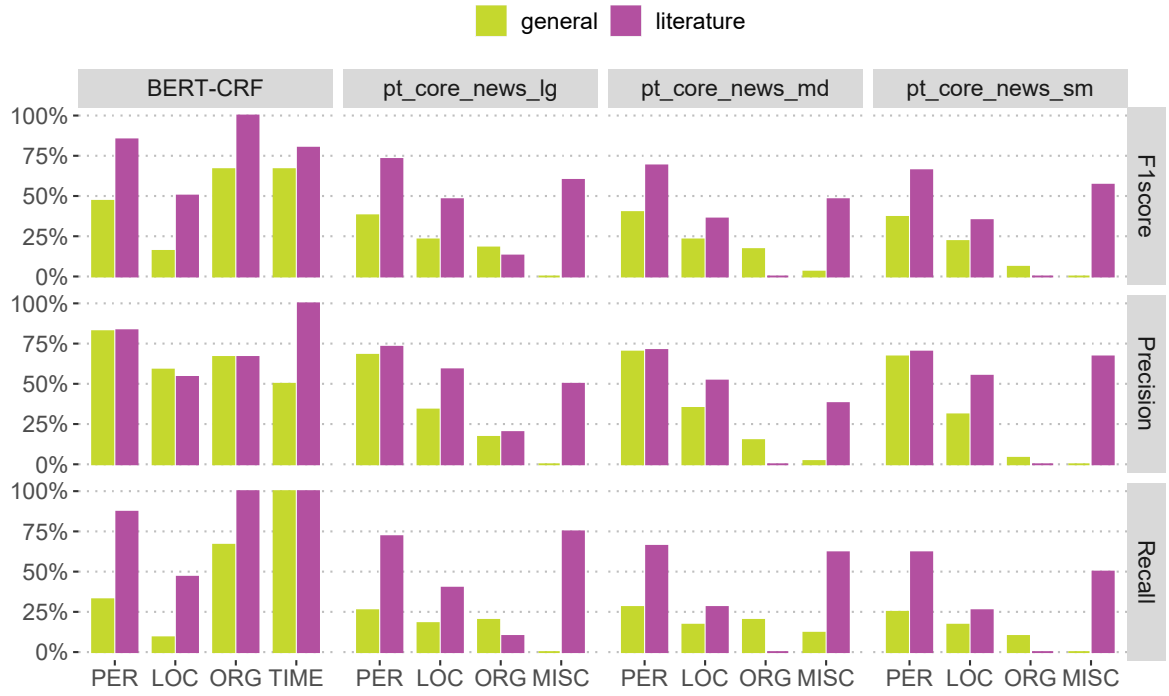

Figure 1: Evaluation metrics by entity categories and domains.

6.3.1. Overall Performance

Our evaluation of spaCy models indicates significant variations in their performance depending on the training data source. When the spaCy models are exclusively trained on the WikiNER dataset, the performance is suboptimal, with all models achieving an F1 Score of less than 35%. In particular, the small-sized spaCy model exhibited the weakest performance (P: 0.44, R: 0.22, F1: 0.29), whereas the large one exhibited the best performance (P: 0.47, R: 0.23, F1: 0.31).

In contrast, the BERT-CRF model trained on the HAREM Selective dataset achieves a high precision (0.79) but a relatively low recall (0.27) and F1 Score (0.41), albeit higher when compared to spaCy models trained on WikiNER. Such a result has already been previously observed, but in 18th-century medical texts written in Portuguese (Zilio et al., 2022). The notable difference in results between spaCy and BERT-CRF can be attributed to various factors, such as the model architecture, training data, and entity category harmonization.

The remarkably low recall prompts further investigation into its underlying causes as it evaluates a model's capability to identify all relevant entities. This consistent low recall can be attributed to various factors: (i) the pre-trained models' training data contains different entity categories and might not align well with the entity categories in our *PPORTAL_ner* corpus (such divergence in entity categories can lead to missed entities when transitioning to a domain-specific dataset); (ii) inherent variability in literary texts, characterized by creative expressions of entities, posing a challenge for models to recognize them consistently; or (iii) literary texts may contain complex hierarchical entities or multi-word expressions that are challenging to capture accurately (e.g., "Capitu's parents" or "Sr. Págua and Dona Fortunata").

Regarding the fine-tuned models, the fine-tuning process on our domain-specific *PPORTAL_ner* corpus led to a substantial and consistent improvement in the performance of all evaluated pre-trained models. The F1 Score for all spaCy models nearly doubled, showcasing the significant positive impact of domain-specific fine-tuning. The BERT-CRF model significantly improved in all three performance metrics, achieving an impressive F1 Score of 0.77 while maintaining high precision and recall.

6.3.2. Entity Analysis

To better detail the performance analysis, Figure 1 depicts the results of the evaluation metrics of each model, categorized by entity type and training domain. Regarding recall, which indicates how well the model is capturing all real entities in the data, the results reveal that across different models on the literature domain, entities falling into the categories of TIME, MISC, and PER yield the highest recall scores, i.e., these entities are well-captured

by the models. Conversely, both LOC and ORG categories show lower recall rates. The limited success in correctly identifying location and organization entities could be attributed to the creative variations in how those categories are referred to or the potential complexities presented by hierarchical or multi-word expressions.

In contrast to the recall metric, the precision scores offer an alternative perspective on the models' performance in correctly identifying entities. Precision indicates the quality of the model's predictions, measuring how much we can trust the entities the model identifies as correct. For all three spaCy models, the entity categories PER and LOC exhibit relatively higher precision rates across the general and literary domains.

The MISC entity category shows a significant improvement when the models are applied to the literary domain, indicating an enhanced ability to correctly identify miscellaneous entities within literary texts. In contrast, the challenge of identifying ORG entities persists across both training domains. Such a challenge can be attributed to the intricacies of recognizing organizational names within literary language, where creative expressions and variations in organization names are common. For instance, within Eça de Queirós's work "Cartas de Inglaterra", references to organizations such as *Mollie Maguire*, a 19th-century Irish secret society, and *the Fenians*, a sister organization of the Irish Republican Brotherhood, are frequently encountered. These organizations may feature non-standard names, posing difficulties for the models in correctly identifying them.

In contrast to the spaCy models, the BERT-CRF model stands out for its overall strong performance, particularly excelling in identifying entities within the ORG category. This better performance may be attributed to the robustness of the BERT-based model, its capacity to capture context and dependencies in the text, and the advantages of pre-training on the HAREM corpus, which aligns more closely with our gold standard.

Finally, the F1 Score combines precision and recall into a single metric that balances the quality of the model's predictions with its ability to identify all true entities. Overall, the F1 Score results highlight the unique challenges posed by different entity categories and the impact of domain-specific training. The spaCy models demonstrate significant improvements in F1 Score when applied to the literary domain, most notably for the MISC category. This suggests that domain adaptation, even with limited specific-domain training data, can positively impact the models' ability to correctly identify entities in literary texts.

However, the ORG category remains a challenge for the spaCy models, with lower F1 Scores across both domains, emphasizing the need for further improvements in identifying these entities in literary contexts. The BERT-CRF model, on the other hand, maintains a higher F1 Score across most entity categories (except for the LOC entity) and both domains, showcasing its superior performance and adaptability to literary NER tasks.

6.3.3. Discussion

Overall, our results underscore the potential for enhancing NER model performance through fine-tuning pre-trained models on domain-specific data, specifically in the context of literary analysis. This improvement is consistent across models of different sizes and architectures, indicating the effectiveness of domain adaptation. Moreover, the variations observed among different entity categories and domains underscore the importance of domain-specific training and the potential benefits of more advanced models, such as BERT-based approaches, for achieving higher accuracy in this unique domain. These findings contribute to a deeper understanding of the challenges and opportunities in NER for literary language.

7. Potential Applications

The *PPORTAL_ner* corpus opens up a plethora of potential applications within the realm of computational literary analysis, offering invaluable resources for advancing research in the digital humanities domain. This section discusses potential applications and provides explicit examples to illustrate how researchers can leverage this corpus for various analytical endeavors.

Literary entity analysis. By extracting and examining named entities such as characters, locations, organizations, and temporal references, researchers can gain deeper insights into literary works' narrative structure, themes, and stylistic elements. For example, researchers can analyze the distribution and frequency of character mentions across different novels to identify recurring motifs or characterize the social dynamics within a literary corpus.

Genre and style classification. By incorporating entity-level features extracted from the *PPORTAL_ner* corpus, researchers can train machine learning algorithms to classify literary texts into different genres or identify the stylistic characteristics associated with specific authors or literary movements. For instance, exploring how certain named entities correlate with the classification of novels into distinct genres.

Cultural and sociopolitical analysis. Exploring the cultural and sociopolitical dimensions embedded within literary texts can provide valuable insights into societal norms, values, and historical

contexts. For example, scholars can investigate how the representation of named entities, such as characters and organizations, reflects broader cultural trends, ideological shifts, or historical events in Portuguese society.

Finally, we believe our *PPORTAL_ner* corpus not only empowers computational literary analysis within the digital humanities domain but also acts as a valuable resource for advancing research in natural language processing (NLP). For example, our corpus can be used as a benchmark dataset for training and evaluating domain-specific named entity recognition (NER) models tailored for Portuguese literary texts (Silva and Moro, 2024a). By training models on the *PPORTAL_ner* corpus, researchers can improve the performance of NER systems in the domain of Portuguese literature.

8. Conclusion

In this paper, we introduced the *PPORTAL_ner* corpus, a novel and comprehensive resource tailored to NER in the domain of Portuguese-language literary texts. Our approach involved a two-step annotation process, where a pre-trained spaCy model provides initial entity suggestions followed by fine-tuning using the Prodigy annotation tool. The resulting corpus addresses the shortage of labeled data for Portuguese literary NER and includes diverse entities commonly found in literary works.

Our experiments demonstrated the significance of using domain-specific data for NER model training. We considered four pre-trained models, including spaCy models of varying sizes and a BERT-CRF model fine-tuned on an in-domain corpus. The results suggest that the fine-tuning process on *PPORTAL_ner* significantly enhances the performance of the pre-trained models. This domain-specific approach is essential for capturing literary texts' distinct characteristics and entities, which can diverge significantly from other domains.

Overall, our corpus can serve as a stepping stone for researchers and developers to explore and develop new applications that require a deeper understanding of entities within literary texts, ultimately enriching the intersection of natural language processing and literary analysis.

Limitations. Despite the relevant insights, some limitations of this work must be acknowledged. First, the corpus considers 25 literary works, which, though diverse, may only partially capture the extensive scope of Portuguese-language literature. Expanding the corpus with additional works from various epochs and styles could enhance its representativeness and applicability across different literary contexts.

Another significant limitation lies in the reliance on a single annotator to correct and refine the pre-annotated data. Although we enforced stringent annotation guidelines to maintain consistency, the subjective nature of entity recognition may introduce minor discrepancies. Moreover, without inter-annotator agreement analysis, it is challenging to assess the consistency of the annotations.

Future work. Several avenues for future work are worth exploring to address the limitations identified and further enhance the utility of the *PPORTAL_ner* corpus. One key aspect involves involving multiple annotators in the annotation process. By incorporating diverse perspectives, we can mitigate the potential biases and errors introduced by a single annotator, thereby improving the corpus's reliability and robustness. Inter-annotator agreement analysis will also provide valuable insights into the consistency and quality of the annotations.

Additionally, while the *PPORTAL_ner* corpus focuses on specific named entity categories such as PER, LOC, GPE, ORG, and DATE, literary texts often contain a plethora of creative entities beyond these categories. These may include events, emotions, and other abstract concepts contributing to literary works' richness and complexity. Future efforts should consider broadening the scope of entity annotation to encompass these unique elements, thereby providing a more comprehensive resource for computational literary analysis.

Finally, we also plan to directly compare our proposed corpus and existing resources, such as the ELTeC-por corpus. Such a comparison will shed light on the strengths and limitations of each corpus, including differences in annotation format, entity categories, and corpus size. Furthermore, this comparative analysis will facilitate the identification of potential synergies and opportunities for collaboration between the two resources, ultimately contributing to the advancement of Portuguese-language computational literary analysis research.

9. Acknowledgements

This work was partially funded by CAPES, CNPq, and FAPEMIG, Brazil.

10. Bibliographical References

Hidelberg Oliveira Albuquerque, Ellen Souza, Carlos Gomes, Matheus Henrique de C. Pinto, Ricardo P. S. Filho, Rosimeire Costa, Vinícius Teixeira de M. Lopes, Nádia Félix F. da Silva, André C. P. L. F. de Carvalho, and Adriano L. I. Oliveira. 2023. Named entity recognition: a survey for the portuguese language. *Proces. del Leng. Natural*, 70:171–185.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in english literature. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 44–54. European Language Resources Association.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2138–2144. Association for Computational Linguistics.

Eckhard Bick. 2006. Functional aspects in portuguese NER. In *Computational Processing of the Portuguese Language, 7th International Workshop, PROPOR 2006, Itatiaia, Brazil, May 13-17, 2006, Proceedings*, volume 3960 of *Lecture Notes in Computer Science*, pages 80–89. Springer.

Niels Dekker, Tobias Kuhn, and Marieke van Erp. 2019. Evaluating named entity recognition tools for extracting social networks from novels. *PeerJ Comput. Sci.*, 5:e189.

Francesca Frontini, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. Named Entity Recognition for Distant Reading in ELTeC. In *CLARIN Annual Conference 2020*, Virtual Event, France.

Abbas Ghaddar and Philippe Langlais. 2017. Winer: A wikipedia annotated corpus for named entity recognition. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 413–422. Asian Federation of Natural Language Processing.

Sara Grilo, Márcia Bolrinha, João Silva, Rui Vaz, and António Branco. 2020. The bdcamões collection of portuguese literary documents: a research resource for digital humanities and language technology. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 849–854. European Language Resources Association.

Anastasios Lamproudis, Aron Henriksson, and Hercules Dalianis. 2022. Evaluating pretraining strategies for clinical BERT models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 410–416. European Language Resources Association.

LDC. 2005. Ace (automatic content extraction) english annotation guidelines for entities. https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-entities-guidelines-v5.6.6.pdf. "Accessed: 2023-10-18".

Igor Morgado, Luis-Gil Moreno-Jiménez, Juan-Manuel Torres-Moreno, and Roseli Suzi Wedemann. 2022. Megalite^{pt}: A corpus of literature in portuguese for NLP. In *Proceedings of the 11th Brazilian Conference on Intelligent Systems*, volume 13654 of *Lecture Notes in Computer Science*, pages 251–265. Springer.

Sean Papay and Sebastian Padó. 2020. Riqua: A corpus of rich quotation annotation for english literary text. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 835–841. European Language Resources Association.

Diana Santos. 2021. Portuguese Novel Corpus (ELTeC-por): April 2021 release.

Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. 2006. HAREM: an advanced NER evaluation contest for portuguese. In *Proceedings of the Fifth International Conference on Language Resources*, pages 1986–1991. European Language Resources Association (ELRA).

Christof Schöch, Roxana Patras, Tomaž Erjavec, and Diana Santos. 2021. Creating the european literary text collection (eltec): Challenges and perspectives. *Modern Languages Open*, 1(25):1–19.

Mariana O. Silva and Mirella M. Moro. 2024a. Evaluating pre-training strategies for literary named entity recognition in Portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, pages 384–393, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Mariana O. Silva and Mirella M. Moro. 2024b. PPORTAL_ner: An Annotated Corpus of Portuguese Literary Entities.

Mariana O. Silva, Clarisse Scofield, Luiza de Melo-Gomes, and Mirella M. Moro. 2022. Cross-collection dataset of public domain portuguese-language works. *J. Inf. Data Manag.*, 13(1).

Mariana O. Silva, Clarisse Scofield, and Mirella M. Moro. 2021. PPORTAL: Public domain Portuguese-language literature Dataset. In *Anais do III Dataset Showcase Workshop*, pages 77–88.

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

Fábio Souza, Rodrigo Frassetto Nogueira, and Roberto de Alencar Lotufo. 2019. Portuguese named entity recognition using BERT-CRF. *CoRR*, abs/1909.10649.

Fábio Souza, Rodrigo Frassetto Nogueira, and Roberto de Alencar Lotufo. 2020. Bertimbau: Pretrained BERT models for brazilian portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems*, volume 12319 of *Lecture Notes in Computer Science*, pages 403–417. Springer.

Sara Stymne and Carin Östman. 2020. Slända: An annotated corpus of narrative and dialogue in swedish literary fiction. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 826–834. European Language Resources Association.

Sara Stymne and Carin Östman. 2022. Slända version 2.0: Improved and extended annotation of narrative and dialogue in swedish literature. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5324–5333. European Language Resources Association.

Krishnapriya Vishnubhotla, Adam Hammond, and Graeme Hirst. 2022. The project dialogism novel corpus: A dataset for quotation attribution in literary texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5838–5848. European Language Resources Association.

Leonardo Zilio, Maria José Finatto, and Renata Vieira. 2022. Named entity recognition applied to portuguese texts from the XVIII century. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese*, volume 3128 of *CEUR Workshop Proceedings*, pages 1–10. CEUR-WS.org.