Evaluating Pre-training Strategies for Literary Named Entity Recognition in Portuguese

Mariana O. Silva
Computer Science Department
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
mariana.santos@dcc.ufmg.br

Mirella M. Moro
Computer Science Department
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
mirella@dcc.ufmg.br

Abstract

In specialized domains, the performance of generic language models can be suboptimal due to significant domain-specific differences. To address such a problem, different pre-training strategies have been proposed for developing domain-specific language models, including cross-domain transfer learning and continuous domain-adaptive pre-training with in-domain data. Within this context, we investigate different pre-training strategies to enhance NER in Portuguese-written Literature. We introduce two models, LitBERT-CRF and LitBERT-Timbau, that leverage domain-specific literary data while building upon general-domain language models. Moreover, we compare cross-domain transfer learning with a general-domain baseline. Overall, our results reveal that both domain-adaptive and transfer learning models outperform the baseline, achieving an F1-Score of over 75% in a strict evaluation scenario and over 80% in a partial scenario.

1 Introduction

Literature, often a reflection of culture and history, is rich in diverse characters, places, and cultural allusions. Named Entity Recognition (NER), as a Natural Language Processing (NLP) task, carries profound importance in such a domain through extracting named entities (Claro et al., 2023). By categorizing essential literary elements, such as character names and locations, researchers can delve into intricate narratives, discerning patterns, tracking character developments, and exploring the sociocultural context within literary works.

Recently, language models based on BERT – Bidirectional Encoder Representations from Transformers (Devlin et al., 2019) have been notably striking in NER tasks. Often combined with bidirectional recurrent networks, attention, and CRF, such models have shown their ability to capture context and relationships, making them particularly well-suited for the complexity of literary entities (Emelyanov and Artemova, 2019; de Menezes Rodrigues et al., 2022). However, the potential of BERT-based models hinges on the availability of labeled data for fine-tuning, a resource that remains somewhat scarce in the context of NER in Portuguese-written Literature.

Literary texts, especially written in Portuguese, present unique challenges due to the intricacies of the language and the richness of cultural allusions (Santos et al., 2022). Furthermore, annotating named entities in literary texts presents its own unique challenges due to the often metaphorical or symbolic use of names, the historical context, and the inherent ambiguity of character roles (Bamman et al., 2019; de Oliveira et al., 2022).

To mitigate the labeled data scarcity issue, language models, pre-trained on extensive unlabeled corpora and fine-tuned on labeled datasets, have revolutionized the landscape of NLP tasks (Qiu et al., 2020; Raffel et al., 2020; Gururangan et al., 2020; Boukkouri et al., 2022). However, these models typically originate from generic, general-domain corpora, such as Wikipedia. In specialized domains like Literature, such an approach may be suboptimal due to the profound disparities in domain-specific terminology, contextual intricacies, and linguistic nuances (Bamman et al., 2019).

In light of these challenges and research gaps, the objective of this work is to investigate and compare different pre-training strategies for BERT-based models designed for the task of NER in Portuguese-written Literature. The main contributions of this paper are as follows:

- We explore and evaluate different pre-training strategies, including cross-domain transfer learning and domain-adaptive pre-training. By comparing such strategies, we provide insights into the most effective approach for enhancing NER in the literary domain.

- We introduce two novel BERT-based models, LitBERT-CRF and LitBERTimbau, specifically tailored for NER in Portuguese-written literature. These models leverage domain-specific literary data while building upon general-domain language models.
- Besides traditional token-based metrics, we perform a thorough evaluation over multiple scenarios and error types. Such analyses include entity-level evaluation metrics that provide a better understanding of the models' performance, including their ability to identify and classify different entity types in literary texts.

2 Related Work

Generic language models can be suboptimal in highly specialized domains due to significant domain-specific differences. Consequently, researchers have introduced pre-training strategies to create domain-specific language models tailored for distinct contexts, including but not limited to clinical applications (Lee et al., 2020), scientific research (Beltagy et al., 2019), financial analysis (Liu et al., 2020), among others (de Menezes Rodrigues et al., 2022).

Domain-specific pre-training often requires in-domain data and can be undertaken through two primary strategies: starting from scratch with a model trained entirely on domain-specific data, or running continuous pre-training of an existing generic language model (Lamproudis and Henriksson, 2022). While the former requires a substantial amount of domain-specific data, computational resources, and time, it can lead to a model that is highly specialized for the target domain.

On the other hand, the domain-adaptive approach consists of the ongoing pre-training of a generic language model by using unlabeled domain-specific text data (Qiu et al., 2020; Rodríguez et al., 2023). Such a strategy is generally more resource-efficient and quicker than pre-training from scratch under both high- and low-resource settings (Gururangan et al., 2020).

Besides domain-specific pre-training, another option is cross-domain transfer learning. Cross-domain transfer learning is an effective strategy, commonly applied when there is limited annotated data in the target domain but a well-pre-trained model from a source domain (Raffel et al., 2020). This approach leverages knowledge transfer from the source to the target domain, leading to models that typically exhibit improved performance and faster training convergence than starting from scratch (Zhuang et al., 2021).

For Named Entity Recognition (NER) tasks, such pre-training strategies play a crucial role. NER, which involves extracting entities like names of people, locations, and organizations from text, can greatly benefit from domain-specific language models (Rodríguez et al., 2023). Such models not only enhance entity recognition but also improve the overall understanding of the context within specialized domains, such as Literature.

Regarding literary works, especially those written in Portuguese, NER models face unique challenges. Literature often reflects culture, history, and diverse characters, producing rich yet complex content. Effectively identifying and categorizing entities in such contexts requires domain-adaptive pre-training strategies (Bamman et al., 2019), allowing models to capture the linguistic nuances specific to the literary domain.

Our study delves into such a domain-specific NER task in Portuguese-written Literature. By investigating various pre-training strategies for NER in specialized contexts, we can uncover which approach is most effective and shed light on the crucial role of domain-specific models in advancing NLP tasks across diverse domains.

3 Methods and Data

This study evaluates two distinct pre-training strategies: (i) domain-specific pre-training and (ii) cross-domain transfer learning. We briefly define both strategies in Section 3.1, as well as describe the pre-training data, configuration, and models in Sections 3.2 to 3.4. Moreover, the fine-tuning process and data are also outlined in Sections 3.5 and 3.6.

3.1 Pre-training Strategies

Language models pre-trained with a specific domain have demonstrated their potential to enhance predictive performance on downstream tasks (Gururangan et al., 2020). Developing domain-specific language models often requires pre-training by using in-domain data through two primary strategies: pre-training from scratch or continuous pre-training of an existing generic language model (Lamproudis and Henriksson, 2022).

Pre-training from scratch entails training a model completely anew, initialized with random weights, on a substantial corpus of in-domain data. This

Table 1: Subset corpus overview.

Features	
Domain	Literary
Languages	PT and PT-Br
# Documents	10
# Tokens	583,788
Size	3MB

Table 2: Hyperparameters used during pre-training.

Hyperparameters	Value
Learning rate	5×10^{-5}
Batch size	16
Max length	512
Epochs	3
MLM probability	15%

process is computationally intensive and often needs extensive resources for successful execution. Therefore, in this work, we focus on the alternative strategy of **domain-adaptive pre-training**, which builds upon pre-existing generic language models.

Another common pre-training strategy is **cross-domain transfer learning**, which involves transferring knowledge from one domain to another. Such a strategy is effective when there is limited annotated data in the target domain but a well-pre-trained model from a related or larger source domain. Transferring knowledge from the source domain to the target domain can often achieve better performance and faster convergence in training compared to training from scratch.

3.2 Pre-training data

For pre-training data, we consider a subset corpus sourced from the PPORTAL dataset (Silva et al., 2021, 2022),[1] an extensive repository of metadata containing over 80,000 public domain literary works in the Portuguese language, predominantly derived from Brazil and Portugal. The subset contains 583,788 tokens from ten literary public domain works. It comprises full-length documents from different authors and literary genres, ensuring high domain diversity and content quality.

To guarantee uniformity and quality of the corpus, each text underwent pre-processing, i.e., removing special characters (excluding hyphens and punctuation marks, given their relevance in literary contexts). Additionally, we have removed any emails and website references. Table 1 shows the main characteristics of the final subset corpus.

3.3 Pre-training setup

All pre-training sessions use the Masked Language Modeling (MLM) as the training task. In this approach, a predetermined percentage of words within a sequence (specifically 15%) are deliberately masked, and the model's primary objective is to predict the identities of these masked words accurately. This task not only sharpens the model's understanding of the language's contextual relationships but also enhances its proficiency in comprehending and generating text.

We set a maximum pre-training duration of three epochs to balance computational resources and time limitations, ensuring that the model could benefit from multiple iterations of pre-training while staying within practical boundaries. Rather than evaluating the pre-training task, each saved checkpoint is evaluated in terms of the performance on a downstream literary NER task.

All models are also pre-trained using the same hyperparameters. Table 2 details which hyperparameters were used during pre-training.

3.4 Pre-training models

We introduce two novel language models for literary Named Entity Recognition in Portuguese. Both models are pre-trained using the MLM task and our subset corpus to incorporate domain-specific data, making the models well-suited for the identification and recognition of named entities in literary texts. We briefly describe each model as follows.

LitBERTimbau. Builds upon the general-domain BERTimbau model (Souza et al., 2020), which initially underwent pre-training with a vast corpus of Portuguese Wikipedia articles. BERTimbau, as a general-domain language model, provides a strong foundation in Portuguese language understanding and general linguistic knowledge.

LitBERT-CRF. Leverages the general-domain BERT-CRF model (Souza et al., 2019), which offers a unique architecture for enhancing Named Entity Recognition (NER). BERT-CRF was initially pre-trained on the brWaC corpus (Filho et al., 2018), a substantial collection of web text in Brazilian Portuguese. It was subsequently fine-tuned on the HAREM dataset (Santos et al., 2006), which

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

contains labeled named entities in Portuguese. The BERT-CRF architecture combines the BERT model with Conditional Random Fields (CRF), a sequence labeling algorithm frequently used for NER tasks.

3.5 Fine-Tuning & Downstream Task

Both pre-trained literary models are fine-tuned on the NER downstream task using a literary annotated corpus (see Section 3.6). We also fine-tuned the general-domain model (BERT-CRF) by using cross-domain transfer learning. That is, we leverage the pre-trained knowledge from a source domain (in this case, general domain) and transfer it to a target domain (in this case, literary domain), allowing the model to adapt to a new domain without starting from scratch (Mou et al., 2016).

The BERT-CRF is fine-tuned using the HAREM corpus (Santos et al., 2006), which includes two different versions. The first version contains a set of ten distinct named entity classes. Here, we consider the other version, called "selective", which focuses on only five classes: Person, Organization, Location, Value, and Time. In alignment with such a selective version, we adjust our annotated corpus by reclassifying GPE entities as LOCATION and DATE entities as TIME.

Throughout the fine-tuning process, all three models are trained for a fixed number of ten 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. 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.

3.6 Fine-tuning Data

For fine-tuning and evaluating the pre-trained models on a downstream task, we consider a dataset manually annotated for literary entities. The corpus is also sourced from the PPORTAL dataset and contains a diverse range of 25 individual literary works.[2] All of these texts were published before 1953, adhering to the current criteria for public domain status in Brazil, with the majority falling within the timeframe spanning from 1554 to 1938. In total, the corpus contains 125,059 tokens, 5,418 sentences, and 5,266 annotated entities.

[2]Note that the subset corpus used during the pre-training comprises different literary works from the 25 selected works for fine-tuning the models.

Table 3: Distribution of entity classes.

Class	Frequency (%)	Examples
PER	3,609 (68.53%)	"Capitu", "the foreigner"
LOC	1,126 (21.38%)	"the village", "the town"
GPE	315 (5.98%)	"Brazil", "Lisbon"
ORG	115 (2.18%)	"the police", "the Church"
DATE	101 (1.92%)	"XVIII century", "1847"

The annotation process was conducted by a single annotator (one of the authors of this paper) and follows a two-step approach involving initial pre-annotation and subsequent correction and refinement using the Prodigy annotation tool.[3] Initially, all 25 literary texts are pre-annotated by using the spaCy model *pt_core_news_lg*. Next, the *ner.correct* recipe in Prodigy is used to refine the gold-standard dataset, considering the *–update* argument to continuously update the model during the annotation loop.

While acknowledging the limitation of a single annotator, future work could explore strategies for multi-annotator involvement, inter-annotator agreement analysis, and the construction of detailed annotation guidelines. Despite the constraints, the single annotator aimed to maintain consistency and accuracy throughout the annotation process. The Prodigy annotation tool facilitated an efficient workflow, allowing for iterative updates to improve annotation quality over successive cycles.

The final corpus contains annotations of PERSON, LOC, GPE, ORG, and DATE entities. 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 glimpse the corpus's content.

4 Experimental Evaluation

This section outlines the experimental evaluation to assess the different pre-training strategies. First, we describe the experimental setup and the evaluation metrics in Sections 4.1 and 4.2. Next, we discuss the results in Section 4.3.

4.1 Experimental Setup

Table 4 shows the main characteristics of each evaluated model. In addition to our primary models (LitBERTimbau and LitBERT-CRF), we evaluate the BERT-CRF model without fine-tuning as a baseline. By comparing the performance of our pre-

[3]https://prodi.gy/

Table 4: Evaluated models overview.

Model	Strategy	Vocab	C_1	C_2
BERT-CRF	Baseline	General	General	General
FT BERT-CRF	Cross-domain transfer learning	General	General	Literary
LitBERT-CRF	Domain-adaptive pre-training	General	Literary	Literary
LitBERTimbau	Domain-adaptive pre-training	General	Literary	Literary

Vocab: Vocabulary | **C_1:** Pre-training corpus | **C_2:** Fine-tuning corpus

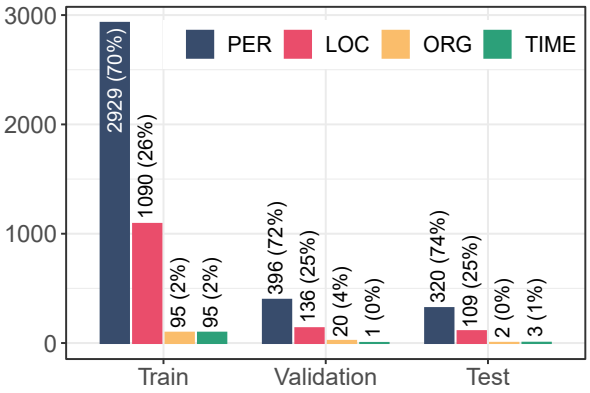

Figure 1: Enter Caption

Table 5: Error types in NER evaluation.

Error type	Description
Correct (C)	True and predicted entities are equal
Incorrect (I)	True and predicted entities do not match
Partial (P)	True and predicted entities are similar
Missing (M)	A true entity that was not predicted
Spurious (S)	A predicted entity that does not exist

trained literary models and the fine-tuned BERT-CRF model (FT BERT-CRF) with the untailored BERT-CRF model, we can validate the impact of our pre-training strategies and evaluate the efficacy of domain adaptation to the specialized domain.

For our experimental setup, we partition the annotated corpus into training, development, and test sets. Such a partition allocates 80% of the sentences to the training set (5,572 sentences), 10% to the validation set (696 sentences), and the remaining 10% to the test set (697 sentences). Figure 1 presents in detail the distribution of entities in each class within the training, validation, and test sets.

4.2 Evaluation Metrics

When assessing NER models, it is a common practice to report metrics at the individual token level. However, this approach may not always be the most comprehensive, especially considering that a named entity can span multiple tokens. To provide a more accurate evaluation, it is essential to account for full-entity accuracy.

Table 6: Evaluation scenarios for NER evaluation.

Eval	Description
Strict	Exact boundary and type matching
Type	Correct entity type assignment regardless of exact boundaries
Partial	Partial boundary matching, regardless of the entity type
Exact	Exact boundary matching, regardless of the entity type

To incorporate the different scenarios into evaluation metrics, we adopt the evaluation schema defined by the *SemEval 2013 - 9.1 task* (Segura-Bedmar et al., 2013), which extends beyond a simple token/tag-based schema. It considers different scenarios, verifying whether all the tokens belonging to a named entity are correctly classified and whether the correct entity type was assigned.

Within such an evaluation schema, five metrics are designed to account for different categories of errors: : Correct (C), Incorrect (I), Partial (P), Missing (M), and Spurious (S). Table 5 provides a description of each error type. Additionally, four distinct evaluation scenarios are considered, examining the models' performance differently: Strict, Type, Partial, and Exact. Table 6 outlines these evaluation scenarios.

For automated evaluation, errors are calculated based on boundary matching, specifically by assessing whether there is an overlap between the true and predicted entities. The overlap is determined by the intersection between the start and end offsets of the true and predicted entities. For instance, if the true entity spans from the third to the seventh token, and the predicted entity spans from the fifth to the ninth token, the overlap would include tokens 5, 6, and 7. This approach enables a nuanced evaluation of partial boundary matching

without imposing rigid percentage constraints.

4.3 Results

Table 7 shows the results of each model, evaluated from the five types of errors and four scenarios. Compared to the baseline (BERT-CRF), all three evaluated models exhibit robust overall performance. BERT-CRF, without fine-tuning, shows a relatively low capacity for capturing entities, reflected in its notably low count of correct entities (C). Additionally, it records a relatively high rate of missing entities (M), implying challenges in capturing certain named entities in the text. Furthermore, it registers some spurious entities (S), indicating a tendency to identify entities that do not exist.

On the other hand, the evaluated models showcase a significant improvement in correctly identifying named entities (C) compared to the baseline. However, they also show a relatively high number of incorrectly classified entities (I), suggesting that such models classify some entities incorrectly. Although they reduced the rate of missing entities (M) compared to the baseline, indicating an improved ability to recognize more named entities, they still face challenges in identifying certain named entities. Notably, they also present many spurious entities (S), suggesting room for fine-tuning to enhance precision and accuracy.

In addition to the error types, we also compute precision, recall, and F1-Score for each scenario (Table 8). Here, precision is the percentage of correctly identified named entities by the model. In contrast, recall represents the model's ability to capture the percentage of named entities in the golden annotations successfully. Such an evaluation is conducted in two distinctive ways, depending on whether an exact match is deemed necessary (for strict and exact scenarios) or if a partial match is acceptable (for partial and type scenarios).

Overall, as detailed next, our results highlight the effectiveness of different pre-training strategies for literary named entity recognition in Portuguese. Both domain-adaptive pre-training and cross-domain transfer learning are valuable approaches for creating language models tailored to this specific NLP task.

Cross-domain transfer learning. The fine-tuned BERT-CRF model (FT BERT-CRF) shows competitive performance. Such a model leverages pre-trained knowledge from a general domain to adapt to the literary domain, significantly capturing entities in the NER task. For the *Strict* scenario, the model presents an F1-Score of 77% with a trade-off between precision and recall. Such balanced performance indicates the model excels in identifying entities correctly and capturing a significant proportion of the named entities.

In the *Exact* scenario, which evaluates exact boundary matching regardless of the entity type, the model also presents a high F1-Score (78%). Such a result highlights its ability to capture a substantial portion of named entities while maintaining precise boundary matching. When considering more relaxed boundary matching scenarios, such as *Type* and *Partial*, FT BERT-CRF outperforms expectations with an F1-Score exceeding 81%. That is, the model can capture a greater proportion of named entities when the boundaries are not exact.

Compared to the other two pre-training strategies, cross-domain transfer learning shows strong results, especially in scenarios where exact boundary matching is not required. The model's competitive results can be attributed to its ability to harness the extensive linguistic and contextual knowledge in general-domain data, thereby expediting its transition into the literary domain.

Domain-adaptive pre-training. Overall, the LitBERT-CRF model outperforms the other models for most evaluation scenarios in Table 8. Nevertheless, its performance closely aligns with the fine-tuned BERT-CRF model. But unlike cross-domain transfer learning, which adapts to the literary domain by leveraging prior knowledge from a general domain, domain-adaptive pre-training directly incorporates domain-specific data into the pre-training process, potentially equipping the model with more specialized linguistic nuances.

The LitBERT-CRF model's strong performance, particularly in the *Strict* scenario, emphasizes its accuracy in precisely identifying literary entities, achieving an F1-Score of 78%. Such a result suggests that the model not only identifies a significant portion of the named entities but also classifies them accurately. The model's consistently high performance extends to other scenarios, especially the *Type* and *Partial*, with F1-Scores above 82%.

In contrast, the LitBERTimbau model, which builds upon the general-domain BERTimbau model, presents competitive yet slightly lower F1 scores across all evaluation scenarios. While it performs well, it falls just short of matching the LitBERT-CRF model's level of accuracy in identi-

Table 7: Evaluation results from the five types of errors and four scenarios. The test set used to evaluate the models has 434 annotated entities.

Model	Strict				Type				Partial				Exact			
	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: Correct | *I:* Incorrect | *M:* Missed | *S:* Spurious | *P:* Partial

Table 8: NER models evaluation results on different training data. The best performance is shown in bold and the second best is underlined.

Model	Strict			Type			Partial			Exact		
	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	<u>0.801</u>	0.279	0.414
FT BERT-CRF	0.770	<u>0.770</u>	<u>0.770</u>	**0.832**	**0.832**	**0.832**	<u>0.817</u>	<u>0.817</u>	<u>0.817</u>	0.784	0.784	<u>0.784</u>
LitBERT-CRF	<u>0.783</u>	**0.774**	**0.779**	**0.832**	0.823	<u>0.827</u>	**0.829**	**0.819**	**0.824**	**0.802**	**0.793**	**0.797**
LitBERTimbau	0.740	0.767	0.753	<u>0.796</u>	<u>0.825</u>	0.810	0.786	0.815	0.800	0.758	<u>0.786</u>	0.771

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

fying literary entities. Such a discrepancy can be attributed to several factors.

First, the initial pre-training of BERTimbau on Portuguese Wikipedia articles might provide a broad linguistic foundation but may not be as tailored to literary nuances as the brWaC and HAREM datasets used by BERT-CRF. Second, the capacity and complexity of the models could vary, with LitBERT-CRF potentially having more parameters or a more sophisticated architecture, which might enhance its entity recognition capabilities.

Entity-level evaluation. Figure 2 shows entity-level evaluation metrics for each model, focusing exclusively on the *Type* scenario. In this specific scenario, a degree of overlap between the boundaries of true and predicted entities is allowed, which adds a layer of flexibility to the evaluation.

Compared to the baseline, both domain-adaptive and cross-domain transfer learning models show high evaluation scores for the PERSON entity class. Although LitBERT-CRF achieves a higher precision (82%), the model exhibits a lower recall rate, which results in a slightly lower F1-Score (85.5%) in comparison to the FT BERT-CRF model (86.3%). Various factors, including differences in the architecture and model capacity, can influence such nuanced differences in NER performance.

The solid overall performance in correctly identifying and categorizing PERSON entities across all models is expected, as such entity class is relatively well-recognized by the generic baseline model. Indeed, PERSON entities often follow common linguistic patterns, making them more accessible to both generic and domain-adaptive language models (Li et al., 2022).

Regarding the other entity classes, the evaluated models present more varied performance results. Specifically for the LOC (location) class, the domain-specific models achieve higher F1 scores compared to the baseline. Such a result suggests that incorporating in-domain knowledge (i.e., literary data) through pre-training strategies significantly improves the extraction of location entities in Portuguese-written Literature.

However, when assessing the ORG (organization) and TIME (time) entity classes, all models face challenges in accurate identification. Despite achieving a relatively high recall rate, indicating their ability to capture a substantial portion of these entities, the precision of the models in recognizing ORG and TIME entities is notably lower. This suggests that while the models successfully capture many instances of organizations and temporal expressions, they also generate numerous false positives, decreasing precision.

The variability in performance across both ORG and TIME entities can be attributed to the complex-

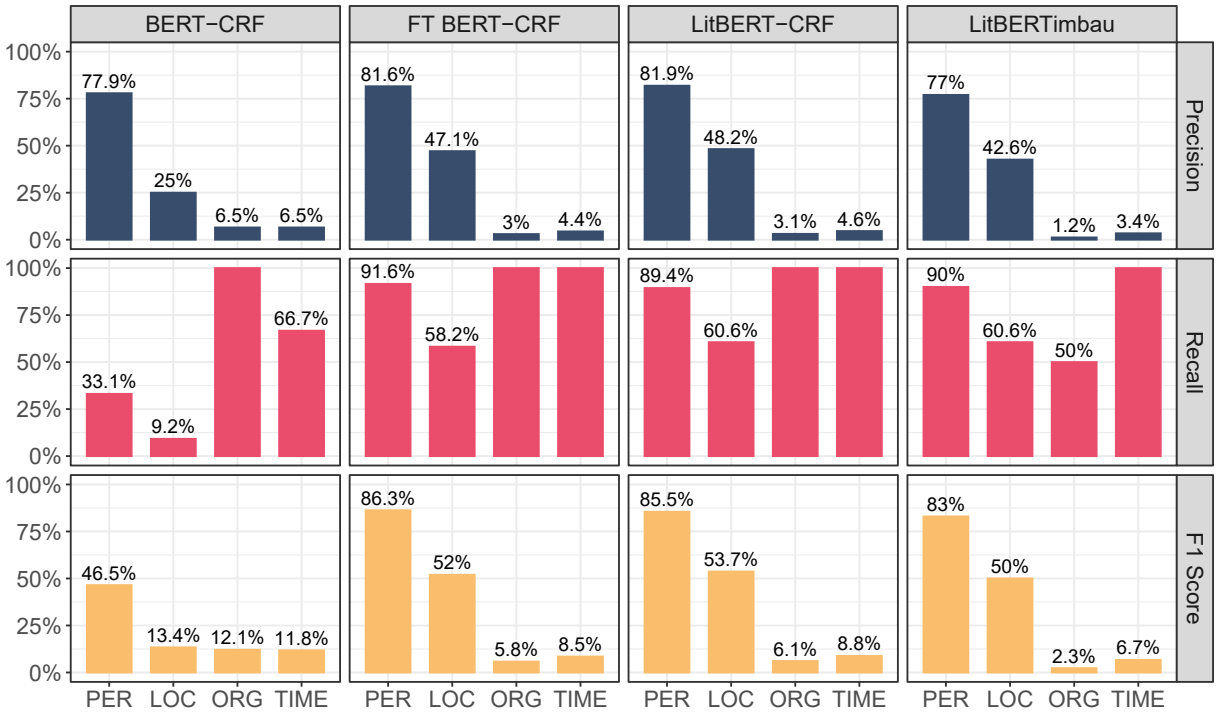

Figure 2: Evaluation metrics for each model, considering the *Type* scenario.

ity and diversity of how organizations and temporal expressions are referenced in literary texts. Authors often employ creative and context-dependent ways of mentioning organizations and time-related information, making it a challenging task for NER models to generalize effectively (Cui and Joe, 2023).

5 Conclusion

In this study, we investigated domain-adaptive pre-training strategies for enhancing Named Entity Recognition (NER) in Portuguese-written Literature. We introduced two domain-adaptive models, LitBERT-CRF and LitBERTimbau, built upon general-domain language models to leverage literary data. Furthermore, we performed a comparative analysis, evaluating cross-domain transfer learning alongside a general-domain baseline. Our findings shed light on the effectiveness of such strategies and their implications for literary NER tasks.

Overall, both domain-adaptive models outperform the baseline BERT-CRF model, showcasing the potential benefits of incorporating domain-specific data into the pre-training process. In particular, LitBERT-CRF outperforms the other evaluated models, with competitive results in different evaluation scenarios, excelling in the strict identification of literary entities.

Moreover, our findings also highlighted the trade-offs associated with different domain-adaptive strategies. The cross-domain transfer learning model (FT BERT-CRF) showed competitive results, especially in evaluation scenarios where exact boundary matching is not required. In contrast, domain-adaptive pre-training models, directly incorporating literary data into the pre-training process, showed superior accuracy in recognizing literary entities.

Our findings open up several avenues for future investigation. For instance, a more extensive and diverse set of literary corpora can be incorporated to capture a broader range of linguistic nuances. Future research can also investigate hyperparameter optimization and advanced training protocols to fine-tune the models more effectively, potentially improving their performance. Finally, while our work focused on the NER task in Portuguese-written Literature, exploring other downstream tasks within the literary domain, such as sentiment analysis, text classification, or even multilingual tasks, can provide insights into the versatility and robustness of the evaluated models.

Acknowledgements

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

References

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.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019*, pages 3613–3618. Association for Computational Linguistics.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, and Pierre Zweigenbaum. 2022. Re-train or train from scratch? comparing pre-training strategies of BERT in the medical domain. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference, LREC 2022*, pages 2626–2633. European Language Resources Association.

Daniela Barreiro Claro, Joaquim Santos, Marlo Souza, Renata Vieira, and Vládia Pinheiro. 2023. Extração de informação. In H. M. Caseli and M. G. V. Nunes, editors, *Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português*, book chapter 17. BPLN.

Shengmin Cui and Inwhee Joe. 2023. A multi-head adjacent attention-based pyramid layered model for nested named entity recognition. *Neural Comput. Appl.*, 35(3):2561–2574.

Rafael Bezerra de Menezes Rodrigues, 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, and Aliel F. Riente. 2022. Petrobert: A domain adaptation language model for oil and gas applications in portuguese. In *Computational Processing of the Portuguese Language - 15th International Conference, PROPOR 2022*, volume 13208, pages 101–109. Springer.

Lucas Ferro Antunes de Oliveira, Adriana S. Pagano, Lucas Emanuel Silva e Oliveira, and Claudia Moro. 2022. Challenges in annotating a treebank of clinical narratives in brazilian portuguese. In *Computational Processing of the Portuguese Language - 15th International Conference, PROPOR 2022*, volume 13208, pages 90–100. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186. Association for Computational Linguistics.

Anton A. Emelyanov and Ekaterina Artemova. 2019. Multilingual named entity recognition using pretrained embeddings, attention mechanism and NCRF. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, BSNLP@ACL 2019*, pages 94–99. Association for Computational Linguistics.

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brwac corpus: A new open resource for brazilian portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*. European Language Resources Association (ELRA).

Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 8342–8360. Association for Computational Linguistics.

Anastasios Lamproudis and Aron Henriksson. 2022. On the impact of the vocabulary for domain-adaptive pretraining of clinical language models. In *Biomedical Engineering Systems and Technologies - 15th International Joint Conference, BIOSTEC 2022*, volume 1814 of *Communications in Computer and Information Science*, pages 315–332. Springer.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.*, 34(1):50–70.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 4513–4519. ijcai.org.

Lili Mou, Zhao Meng, Rui Yan, Ge Li, Yan Xu, Lu Zhang, and Zhi Jin. 2016. How transferable are neural networks in NLP applications? pages 479–489.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *CoRR*, abs/2003.08271.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Dalia Andrea Rodríguez, Julia Diaz-Escobar, Arnoldo Díaz-Ramírez, and Leonardo Trujillo. 2023. Domain-adaptive pre-training on a BERT model for the automatic detection of misogynistic tweets in spanish. *Soc. Netw. Anal. Min.*, 13(1):126.

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 and Evaluation, LREC 2006*, pages 1986–1991. European Language Resources Association (ELRA).

Diana Santos, Roberto Willrich, Marcia Langfeldt, Ricardo Gaiotto de Moraes, Cristina Mota, Emanoel Pires, Rebeca Schumacher Fuão, and Paulo Silva Pereira. 2022. Identifying literary characters in portuguese - challenges of an international shared task. In *Computational Processing of the Portuguese Language - 15th International Conference, PROPOR 2022*, volume 13208, pages 413–419. Springer.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero-Zazo. 2013. Semeval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2013*, pages 341–350. The Association for Computer Linguistics.

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.

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 *Intelligent Systems - 9th Brazilian Conference, BRACIS 2020*, volume 12319 of *Lecture Notes in Computer Science*, pages 403–417. Springer.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A comprehensive survey on transfer learning. *Proc. IEEE*, 109(1):43–76.