Procesamiento del Lenguaje Natural, Revista nº 70, marzo de 2023, pp. 171-185 recibido 16-12-2022 revisado 31-01-2023 aceptado 09-02-2023

Named Entity Recognition: a Survey for the Portuguese Language

Reconocimiento de Entidades Nombradas: una investigación para el idioma Portugués

Hidelberg O. Albuquerque[1,2], Ellen Souza[1,3], Carlos Gomes[4], Matheus Henrique de C. Pinto[3], Ricardo P. S. Filho[5], Rosimeire Costa[5], Vinícius Teixeira de M. Lopes[6], Nádia F. F. da Silva[3,5], André C. P. L. F. de Carvalho[3], Adriano L. I. Oliveira[2]

[1]MiningBR Research Group, Federal Rural University of Pernambuco, Brazil
[2]Centre of Informatics, Federal University of Pernambuco, Brazil
[3]Institute of Mathematics and Computer Sciences, University of São Paulo, Brazil
[4]Institute of Mathematics and Technology, Federal University of Catalão
[5]Institute of Informatics, Federal University of Goiás, Brazil
[6]Federal University of Campina Grande
{hidelberg.albuquerque,ellen.ramos}@ufrpe.br, alio@cin.ufpe.br,
andre@icmc.usp.br, matheuscerqueira@usp.br, cadyoba@gmail.com,
{rpsfilho93,rosimeire_pereira}@discente.ufg.br, nadia.felix@ufg.br,
vinicius.teixeira.melo.lopes@ccc.ufcg.edu.br

Abstract: Named Entity Recognition (NER) is an important task in Natural Language Processing, as it is a key information extraction sub-task with numerous applications, such as information retrieval and machine learning. However, resources are still scarce for some languages, as it is the case of Portuguese. Thus, the objective of this research is to map NER techniques, methods and resources for the Portuguese language. Manual and automated searches were applied, retrieving 447 primary studies, of which 45 were included in our review. The growing number of studies reveal a greater interest of researchers in the area. 21 studies focused on the comparative analysis between techniques and tools. 24 new or updated NER corpora were mapped, in several domains. The most used text pre-processing techniques were tokenization, embeddings, and PoS Tagging, while the most used methods/algorithms were based on BiLSTM, CRF, and BERT models. The most relevant researchers, institutions and countries were also mapped, as well as the evolution of publications.
Keywords: Named Entities Recognition, Review, Portuguese.

Resumen: El Reconocimiento de Entidades con Nombre (en inglés, *NER*) es una tarea importante en el Procesamiento del Lenguaje Natural, ya que es una subtarea clave de extracción de información con numerosas aplicaciones, como la recuperación de información y el aprendizaje automático. Sin embargo, los recursos aún son escasos para algunos idiomas, como es el caso del portugués. Por lo tanto, el objetivo de esta investigación es mapear técnicas, métodos y recursos de NER para la lengua portuguesa. Se aplicaron búsquedas manuales y automatizadas, recuperando 447 estudios primarios, de los cuales 45 se incluyeron en nuestra revisión. El creciente número de estudios revela un mayor interés de los investigadores en el área. 21 estudios se centraron en el análisis comparativo entre técnicas y herramientas. Se mapearon 24 corpora NER nuevos o actualizados, en varios dominios. Las técnicas de preprocesamiento de texto más utilizadas fueron *tokenization*, *embeddings* y *PoS Tagging*, mientras que los métodos/algoritmos más utilizados fueron los basados en *BiLSTM*, *CRF* y de los modelos *BERT*. También se mapearon los investigadores, instituciones y países más relevantes, así como la evolución de las publicaciones.
Palabras clave: Reconocimiento de Entidades Nombradas, Revisión, Portugués.

ISSN 1135-5948 DOI 10.26342/2023-70-14

1 Introduction

Natural Language Processing (NLP) is one of the multidisciplinary areas involving the fields of Linguistics and Artificial Intelligence. NLP is a challenge for researchers and professionals because it corresponds to how natural language, with all its richness, complexities, and variances can be transformed and used by computational systems (Finatto, Lopes, and Silva, 2015). Named Entity Recognition (NER) is a NLP technique that aims to identify entities in the text and classify them into sets of universal syntactic or semantic categories (Maynard, Bontcheva, and Augenstein, 2016), or the ones specific to a particular language or domain (De Araujo et al., 2020). The classified data and the extracted features are used in text mining systems (Nadeau and Sekine, 2007) or in Machine Learning models (Bonifacio et al., 2020), and other applications.

For the recovery, processing and textual analysis to be effective, it is important to determine which methods are the best for each domain or language, which can explain why much of the research focuses on a monolinguistic approach (Akbik et al., 2016). Studies about NER for Portuguese show evidence that the models used for this language have challenges not found for other languages, which can be explained by the low volume of corpora, tools and pre-trained models developed for Portuguese (Castro, 2018). Researches in this language need a greater effort, mainly in the development of resources, approaches and tools, as occurs to English language (Pirovani, 2019).

Based on the guidelines proposed by Kitchenham, Charters, and others (2007) and Petersen et al. (2008), the main objective of this work is to characterize the current researches that report the use of techniques for NER in Portuguese, seeking to answer the general research question: *What is the current status of NER tasks for the Portuguese language?*

In this way, the automated and manual search procedures retrieved 447 papers published between January/2010 and June/2022 from which 63 were pre-selected and 45 were included in this study. Data extracted from primary studies were systematically structured and analyzed to answer historical, descriptive, and classificatory research questions presented below:

- RQ1: What are the existing corpora for NER in Portuguese Language?
- RQ2: What algorithms, techniques, and tools were used to build and validate the Portuguese NER models?
- RQ3: How the Portuguese NER models have been used in NLP tasks?
- RQ4: What has been the evolution of the number of publications until the year 2022?
- RQ5: What individuals, organizations and countries are the main contributors in this research area?

The remainder of this paper is structured as follows: Section 2 presents the related work. Section 3 details the review method. In Section 4, a comprehensive set of results is presented. Section 5 discusses the results, and contains conclusions and directions for future works. Finally, the Appendix A presents the list of primary studies selected, with their respective access links.

2 Related Work

The mapping of techniques, methods and resources are an indisputable key to the progress of any research, and an invaluable source for any researcher. This section highlights some initiatives with relevant contributions.

Nadeau and Sekine (2007) performed a survey on Named Entity Recognition and Classification (NERC), in a hundred studies published in English in ten different events, between 1991 and 2006. The review reports studies performed in over 20 languages, a wide range of named entity types, their semantic challenges, and hierarchical subcategories. Most studies have focused on limited domains and textual genres. The work also provides an overview of the studies selected from the challenges or techniques used: diversity of languages, types of text used, textual genres, application domains, corpora, disambiguation rules, machine learning algorithms, features, as well as evaluation methods. Finally, the work highlights the great importance of NER for NLP-based systems.

Sun et al. (2018) conducted a research using 162 publications from NLP conferences, between the years 1996 to 2017. The authors discuss about two aspects of research in NER: the first one, based on target languages (covering papers from more than 200 languages, with mono, bi, and multilingual approaches), and the second one, a more technical approach with statistical analysis used in NER tasks. Some results brought by the authors were the mapping of the number of publications, the most used languages, the proportion between publications with different approaches, and the different methods.

Yadav and Bethard (2019) explore the advances of recent architectures with better deep learning results for state-of-the-art. Studies that combined learning models based on minimal resources were selected, which were compared with models of feature-based learning and with different representations of words. An automatic search was used in three search engines, with a search string. The papers were initially classified by total of citations, being pre-selected those that used an unpublished NER neural architecture, or a representation of a high-performance model for NER datasets, independent of domain or language. When published architecture was found, citation tracing back to the architecture's original source was performed. 154 papers were reviewed and 83 were selected. This results were subdivided into NER datasets, evaluation metrics and systems based on different techniques and architectures. The authors compared the results found in four languages (Spanish, Dutch, English and German), highlighting the need for future progress using insights from previous work applied to current neural network models.

Li et al. (2020) also focuses on studies of deep learning models for NER. After a brief review of traditional NER techniques, the authors make an intense review of studies, applications and deep learning techniques for NER, using universal entities in English. It was proposed a new taxonomy, which systematically organizes the approaches along three axes: distributed representations for input, context encoder (to capture contextual dependencies), and tag decoder (to predict word labels). In addition, the paper presents relevant secondary results, such as, corpora annotated in English, NER tools, summary of recent works on neural NER, besides presenting challenges and future directions.

3 Review Method

Secondary studies review all the primary studies relating to a specific research question with the aim of integrating/synthesizing evidence related to a subject (Keele and others, 2007). In this study, the search for primary studies was done in six steps, as shown in Fig. 1, detailed in the following subsections.

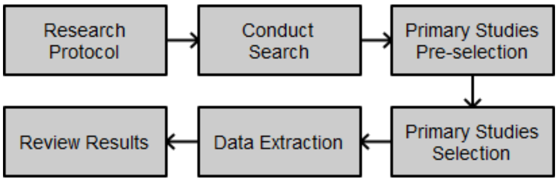

Figure 1: Review method (Keele and others, 2007).

3.1 Research Protocol

The research protocol outcome is the review scope, which includes, among other things, the research questions presented in Section 1, the inclusion and exclusion criteria, and data sources selection. Primary studies that reported NER corpora, algorithms, methods, and techniques for Portuguese were searched in the literature, in the last 12 years. Studies that met at least one of the following exclusion criteria were removed from this review: (i) written in a language other than English or Portuguese; (ii) not available on online scientific libraries; (iii) keynote speeches, workshop reports, books, theses and dissertations.

3.2 Conduct Search

Two different types of searches were performed: automated and manual. First, a manual search was performed, followed by the automatic one, performing the removal of duplicate studies. In the former, a search string was used to retrieve papers from digital libraries, using the following terms and their synonyms:

- Named Entity Recognition: NER, Recognition of Entities, *Reconhecimento de Entidades Nomeadas*, *REN*, *Reconhecimento de Entidades Mencionadas*, *REM*, *Reconhecimento de Entidades com Nome*, Entity Extraction Task, Name Entity, NE.
- Portuguese: Portuguese, *Língua Portuguesa*, *Português*.

ACL Anthology[1] and Google Scholar[2] digital libraries were selected to conduct

[1] https://aclanthology.org

[2] https://scholar.google.com

the automated searches, covering the period between January/2010 and June/2022. The ACL Anthology currently hosts 80,558 papers on the study of computational linguistics and natural language processing, indexing several events and journals in the NLP area.

For the manual search, we selected the main venues focusing on the Portuguese language (Souza et al., 2016; Souza et al., 2018): The International Conference on the Computational Processing of Portuguese (PROPOR); Portuguese Conference on Artificial Intelligence (EPIA); Brazilian Symposium on Information and Human Language Technology (STIL); Brazilian Conference on Intelligent Systems (BRACIS); Brazilian National Meeting of Artificial and Computational Intelligence (ENIAC) and Language Resources and Evaluation Conference (LREC), which is the major event on Language Resources and Evaluation for Language Technologies in several languages.

Manual Search	**Pre-sel.**	**Sel.**
BRACIS	7	6
ENIAC	3	3
EPIA	2	1
LREC	7	4
PROPOR	3	3
STIL	7	5
TOTAL	**29**	**22**
Automated Search	**Pre-sel.**	**Sel.**
EPIA	2	1
IberLEF	4	4
LREC	5	3
Events with one paper	17	9
Journals with one paper	6	6
TOTAL	**34**	**23**
Manual+Automated	**63**	**45**

Table 1: Review by data sources.

3.3 Primary Studies: Pre-selection and Selection

A pre-selection was accomplished in accordance with the inclusion and exclusion criteria established in the review protocol. Primary studies were analyzed using the same procedure for both automated and manual search strategies. Two researchers applied the inclusion and exclusion criteria, after reading the title, abstract, and keywords.

The selection process was done by six researchers. Each researcher was responsible for reading the *full* paper and presenting it at a weekly consensus meeting, where all researchers decided to include or exclude the primary studies. Table 1 shows the number of potentially relevant primary studies (Preselection step) and number of included studies (Selection step).

3.4 Data Extraction and Review Results

In this fifth step, data from included primary studies were extracted and synthetized to answer the research questions. The researchers worked independently to extract data from the included papers, using an extraction form. Finally, one researcher inspected the extracted data and *ad hoc* consensus online meetings were held. The details of the extraction form with its results were packed for later replication[3].

4 Results

In this section, the obtained results are presented, organized according to the five specific research questions.

4.1 RQ1: What are the existing corpora for NER in Portuguese Language?

Information about corpora, entities, annotation method, agreement level, domain, type of text used, and language variants were extracted from primary studies (PS).

In general, the studies presented a pattern in the use of manual annotation, absence of agreement level measure among annotators, use of formal texts in the construction of the corpora, and absence of comparison of entities for corpora with the same domain. The works that differ are: (i) PS05, PS07, PS21 and PS22, which used automatic annotation; PS06, PS17, PS25, PS34 and PS45 used hybrid annotation, and PS43 does not inform the used method; (ii) for the agreement measure, PS19 presented a measure of 95.8 % (without specifying which one was used) and PS01, in general, 91 % (using Cohen's Kappa); (iii) regarding the type of text, PS18 do not inform which type was used, PS09 used a mix between formal and informal usage. PS15, PS16, PS22, PS42, PS43, and PS45 used informal texts; (iv) regarding the difference of entities in the same domain, PS24 used only two entities against six of

[3] Available in https://bit.ly/extraction-form-survey

Corpora/PS	Entities	Annotation method	Domain	Text type	Language variant
Aposentadoria/PS25	Act, Act_Name, Class, Cod_Enrollment_Act, Company_Act, Legal_Fund, Position, Frame, Pattern, and Process.	Hybrid	Legal	Formal	PT-BR
DataSense NER Corpus/PS11	Bank Identification Number, Credit Card Number, Date, Driving License Number, E-mail address, Identification Number, Job, Local, Med, National Health Number, Organization, Passport Number, Person, Postal Code, Social Security Number, Tax Identification Number, Telephone Number, and Value.	Manual	Sensitive Data	Formal	PT-EU
Dicionário Histórico-Biográfico Brasileiro (DHBB)/PS17	Document, Event, Local, Organization, Person, Political Formulation, and Time.	Hybrid	History	Formal	PT-BR
DrugSeizures-Br/PS07	Drug, Location, Organization, Other, Person, and Time.	Automatic	Legal	Formal	PT-BR
EHR-Names/PS40	Person	Manual	Medical	Formal	PT-BR
Financial Market Corpus/PS34	Organization, Person and Place	Hybrid	Financial	Formal	PT-BR
GeoCorpus/PS03 and GeoCorpus-2/PS10	Aeon, Era, Period, Epoc, Age, Siliciclastic Sedimentary Rock, Carbonate Sedimentary Rock, Chemical Sedimentary Rock, Organic-rich Sedimentary Rock, Brazilian Sedimentary Basin, Basin Geological Context, Lithostratigraphic Unit, and Miscellaneous.	Manual	Geology	Formal	PT-BR
LeNER-Br/PS20	Legal cases, Legislation, Location, Organization, Person, and Time.	Manual	Legal	Formal	PT-BR
PS06	CPF_CNPJ, Marital status, Name, Nationality, OAB, and RG.	Hybrid	Legal	Formal	PT-BR
PS16	Date, Location, Organization and Person	Manual	General	Informal	PT-EU
PS19	Characterization, Test, Evolution, Genetics, Anatomical Site, Negation, Additional Observations, Condition, Results, DateTime, Therapeutics, Value, and Route of Administration.	Manual	Neurology	Formal	PT-EU
PS22	Location, Organization, and Person	Automatic	Journalistic	Informal	PT-EU
PS23	Location, Organization, and Person.	Manual	Police	Formal	PT-BR
PS24	Legal cases and Legislation	Manual	Legal	Formal	PT-BR
PS35	Person	Manual	Legal	Formal	PT-BR
PS36	Place and Person	Manual	Literature	Formal	PT-BR
PS42	Brand, Camera quality, Color, Display size, Internal memory, Model, Operating system, Processor, SIM card capacity, and WIT (What Is This)	Manual	E-commerce	Informal	PT-BR
PS43	Organization, Person, and Location	*Not Informed*	Jornalistic	Informal	PT-BR
PS45	Location and Event	Hybrid	Traffic	Informal	PT-BR
Second HAREM/PS15	Abstraction, Event, Location, Organization, Other, Person, Thing, Time, Title, and Value	Manual	General	Informal	PT-BR & PT-EU
SESAME/PS21	Location, Organization, and Person	Automatic	General	Formal	PT-BR
Summ-it++/PS05	Abstraction, Event, Organization, Other, Person, Place, Thing, Time, Value, and Work.	Automatic	General	Formal	PT-BR
UlyssesNER-Br/PS01	Date, Event, Law Fundation, Law product, Location, Organization, and Person	Manual	Legislative	Formal	PT-BR

Table 2: Portuguese NER corpora.

PS20; PS37 compares universal entities, such as Person, Place, Organization, Value, Time, Abstraction, Work, Event, and Thing using HAREM and SPA Conll-2002 models; PS01 and PS11 adopt more specificity when compared to the entities of PS15 and PS20. In turn, PS15 updates the golden collection of HAREM (Santos and Cardoso, 2006), presenting the main improvements: removal of repeated texts, cleaning of uncertain sequences, accounting of partially correct entities and systematization in the treatment of entities. HAREM is a huge Portuguese language NER corpus widely used by the Portuguese NLP community.

Among the primary studies that were included, 27 (60 %) did not indicate whether or how interference occurs based on text ty-

Domain	**Corpora/PS**	**Public link**
E-commerce	PS42	—
Financial	Financial Market Corpus/PS34	http://bit.ly/finmktcorpus
General	brWaC/PS07, PS38, and PS39	https://bit.ly/BrWaC-corpus
	Floresta Sintática/PS11	https://bit.ly/floresta-corpus
	Freeling/PS04 and PS12	https://bit.ly/freeling-corpus
	HAREM I, HAREM II and MiniHAREM/PS15	http://bit.ly/haremcorpus
	Paramopama/PS09 and PS21	https://bit.ly/paramopama
	PS16	https://bit.ly/ps16-ptools
	SESAME/PS21	https://bit.ly/sesamecorpus
	Summ-it++/PS05	https://bit.ly/summ-it
	WikiNER/PS08, PS09, PS21, PS23, and PS30	https://bit.ly/wikiner
Geology	GeoCorpus/PS03	https://bit.ly/GeoCorpus
	GeoCorpus-2/PS10	https://bit.ly/geocorpus2
History	Dicionário Histórico-Biográfico Brasileiro (DHBB)/PS17	https://bit.ly/DHBB-corpus
Jornalistic	aTribuna/PS08 and PS30	https://bit.ly/atribuna-corpus
	CETEMPúblico/PS43	https://bit.ly/CETEMPublico
	CETENFolha/PS43	https://bit.ly/cetenfolha
	PS22	—
	PS43	—
	SIGARRA News/PS16	https://bit.ly/sigarranews
Legal	Acordaos-TCU/PS07	https://bit.ly/acordaos-tcu
	Aposentadoria/PS25	https://bit.ly/aposentadoria-corpus
	Data-lawyer/PS09	—
	DrugSeizures-Br/PS07	—
	LeNER-Br/PS20	https://bit.ly/lener-br
	PS06	—
	PS24	—
	PS35	—
Legislative	UlyssesNER-Br/PS01	https://bit.ly/ulyssesner-br
Literature	PS36	—
Medical/ Clinical	EHR-Names/PS40	—
	PS19	—
	PS44	https://bit.ly/ps44-BioBERTpt
	SemClinBr/PS44	https://bit.ly/SemClinBr
Police	PS23	—
Sensitive Data	DataSense NER Corpus/PS11	—
Traffic	PS45	—

Table 3: Corpora per domain.

pe or domain in the NER task. Analyzing the studies that provide this information, it is possible to correlate domain specificity to decreased efficiency in the results.

PS21 indicates that the general domain facilitates information extraction and enables cross-referencing of information in order to increase complexity. In the legal domain, PS20 presents unique entities to represent laws and legal cases, PS35 reports that capitalization of proper names increases the generation of false positives, and PS01 is the only study that present entities for the legislative subdomain. In PS10 and PS17, for the Geology and History domains, respectively, the universal entities (such as Person, Organization, and Place) were insufficient to represent the complexity of the research.

In the medical/clinical domain, there is a significant worsening of F-measure compared to the general domain in PS40, caused by the concatenation of the corpora used; in PS41, this worsening is justified by the use of terms and abbreviations unique to the domain, while in PS44, the existence of multiple labels is pointed out as the main factor. PS08, PS09, PS30, and PS32 show that, in comparison, the NER task performs poorly for the clinical domain and reasonably to the police domain when compared to the general domain. In the E-commerce domain, PS42, the lack of syntactic structure makes attribute extraction difficult. Finally, PS43 states that in the journalistic domain, when approaching the gene-

Techniques	Total	Primary Studies
Tokenization	19	PS02, PS06, PS09, PS11, PS19, PS20, PS21, PS25, PS27, PS28, PS31, PS32, PS33, PS35, PS36, PS37, PS39, PS42, PS44
Embeddings	13	PS01, PS06, PS08, PS09, PS10, PS19, PS20, PS24, PS33, PS38, PS39, PS40, PS41
PoS Tagging	10	PS03, PS04, PS05, PS11, PS12, PS14, PS19, PS22, PS28, PS36
Lower case	6	PS20, PS33, PS35, PS37, PS39, PS42
Spell-check	5	PS02, PS16, PS29, PS32, PS34
Format Conversion	5	PS10, PS13, PS14, PS24, PS28
Special character removal	5	PS34, PS35, PS39, PS42, PS45
Features	4	PS02, PS12, PS13, PS32
Non-text removal	4	PS03, PS21, PS39, PS45
Removal of repeated sentences (outliers)	3	PS01, PS10, PS42
Removal of pre-textual or post-textual elements	3	PS03, PS21, PS36
Stop Words	2	PS27, PS42
Removal of HTML Tags	2	PS32, PS39

Table 4: Pre-processing techniques.

ral domain, the task is facilitated, probably due to language simplification.

Few articles presented applications for different language variants. PS12, PS13, PS25, and PS33 do not report whether language variation interferes with the NER tasks. The strategy of PS21 (using the DBPedia ontology[4]), avoids the inclusion of Portuguese language papers. PS42 indicates that, when tuning the BERTimbau model[5] with HAREM, there was a subtle worsening in performance. PS27, when comparing Portuguese variants, registered a difference in performance, justified by linguistic and cultural differences.

From the selected works, 21 (∼ 47 %) studies did not create, modify or update corpora: PS02, PS04, PS08, PS09, PS12 to PS14, PS18, PS26 to PS33, PS37 to PS39, and PS41. These works carried out comparative analyzes between tools, methods or the application of NER tasks in multidomains, in different textual genres or in textual semantic relations, using some variation of the HAREM corpus. Table 2 summarizes the 24 studies in which a new corpus was created or updated from an existing corpus. Some of them did not have a clearly identifiable name for the created or modified corpus, and were instead referred to by the PS number. Finally, Table 3 presents all corpora used in the studies, organized by domain. Some of these corpora are publicly available. Due to limited space, the entities labels are detailed in the document specified in the third footnote (Section 3.4).

4.2 RQ2: What algorithms, techniques, and tools were used to build and validate the Portuguese NER models?

37 PS (∼82 %) mention the use of some type of preprocessing techniques as shown in Table 4. The most used techniques were tokenization, embeddings, and PoS Tagging. Regarding pre-processing, it was possible to observe that some works focused on the use and analysis of the influence of embeddings and features, among others.

Table 5 presents the most used methods, algorithms, and tools: learning models based on BiLSTM were the most used, followed by more traditional methods that use only CRF, as well as systems developed specifically for NER tasks, such as the NERP-CRF and PALAVRAS parser. More recent works used BERT deep learning model. The metrics used in most studies were Accuracy, Precision, Recall, and F-score. However, as not all papers presented all this information, the table shows the F-score range of performance obtained. Other tools, algorithms or techniques that had only one mention were not listed.

[4] https://www.dbpedia.org/resources/ontology

[5] https://github.com/neuralmind-ai/portuguese-bert

Algorithms/Methods/Tools	Total	F-score(%)	Primary Studies
BiLSTM+CRF	10	67 ∼ 97	PS06, PS09, PS10, PS20, PS24, PS33, PS38, PS39, PS40, PS41
CRF	7	48 ∼ 97	PS01, PS04, PS13, PS19, PS28, PS36, PS45
CRF+LG	6	53 ∼ 76	PS02, PS09, PS29, PS30, PS31, PS32
BiLSTM	3	53 ∼ 93	PS01, PS11, PS43
NERP+CRF	3	53 ∼ 93	PS05, PS12, PS14
PALAVRAS	3	57 ∼ 62	PS04, PS12, PS17
BERT+CRF	2	75 ∼ 95	PS34, PS44
FreeLing	2	54 ∼ 56	PS04, PS12

Table 5: Algorithms/methods/tools.

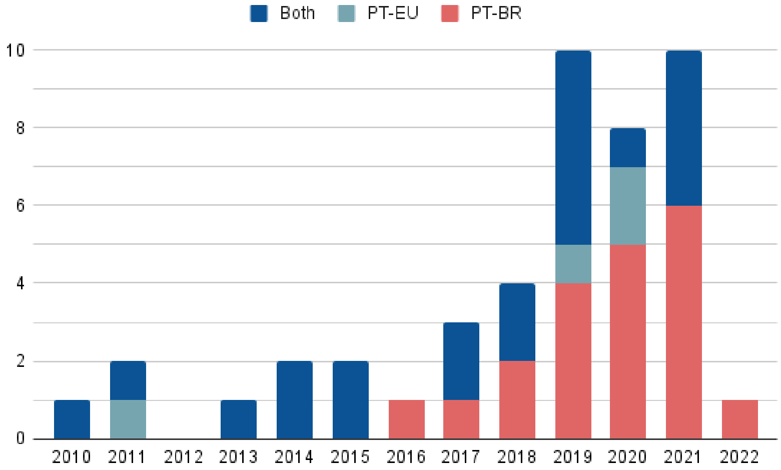

Figure 2: Temporal distribution of PS, between January/2010 to June/2022[6].

4.3 RQ3: How the Portuguese NER models have been used in NLP tasks?

As previously mentioned, a significant part of the selected studies focused their research on comparative analyzes between NER tasks *per se*. Some of these works showed great potential for applicability and/or improvement of information extraction systems, such as studies PS12, PS13, PS20, PS23, and PS29.

Among the selected studies, 20 applied NER models directly or indirectly in other NLP tasks, which are shown in Table 6. Of these, Information Retrieval appears as the main applied task (55 %), followed by Relation Extraction (30 %), Morphosyntactic Annotation and Semantic Similarity tasks (both with 15 %). Two studies explored the impact of language models with Machine Learning, using non-common models in the state-of-the-art, proposing improvements based on the results found.

4.4 RQ4: What has been the evolution of the number of publications until the year 2022?

As shown in Figure 2, the first primary study selected was published in 2010 (PS15). The number of studies ranged in the first years (1∼2 articles per year), increasing from 2017, with the exception of 2022 (the current year of this research[6]). Primary studies were classified according to the Portuguese language variant: studies that apply only European Portuguese represent 8.89 %, while Brazilian Portuguese comprises 44.44 %. Finally, 46.67 % of studies used text written in both languages, most of them with a HAREM corpus variation.

[6] The searches in automated sources were performed in June/2022, which explains the low number of publications for this year.

Task	Total	Subtask	Total	Primary Studies
Information Retrieval	12	-	-	PS01, PS07, PS10, PS11, PS15, PS17, PS18, PS19, PS22, PS42, PS44, PS45
Relation extraction	6	-	-	PS05, PS09, PS15, PS17, PS26, PS34
Morphosyntactic Annotation	3	-	-	PS05, PS17, PS22
Semantic similarity	3	-	-	PS30, PS38, PS44
Classification	2	Document Classification	1	PS18
		Text Classification	1	PS42
Co-reference resolution	2	-	-	PS05, PS15
Data Privacy	2	De-Identification	1	PS40
		Sensitive Data	1	PS11
Machine Learning for NLP	2	Deep Active Learning	1	PS25
		Transfer Learning	1	PS07
Tracking	1	-	-	PS45
Word sense disambiguation	1	-	-	PS44

Table 6: Related NLP tasks and subtasks.

4.5 RQ5: What individuals, organizations and countries are the main contributors in this research area?

A total of 145 researchers from 45 organizations were mapped. The data showed that Brazil has a greater number of researchers (∼79 %) and research institutes (∼67 %) in the area. Tables 7 and 8 list the main researchers and institutions, with emphasis on the Pontifical Catholic University of Rio Grande do Sul (PUCRS) and researchers Renata Vieira, Daniela O. F. do Amaral and Joaquim Santos, from the same institution. The vast majority of institutions in the papers are Colleges, Universities or Institutes of Higher Education (∼71 %). We also believe it is worth mentioning some private institutions that provided support for research in NER, such as Petrobras Research and Development Center (PS10), IBM Research (PS17, PS37, and PS39), Americanas S.A. Digital Lab (PS42), Viatecla SA (PS22), and some institutions linked to public or political administration, like Public Ministry of the State of Mato Grosso do Sul (PS07), Brazilian Federal Police (PS18 and PS23), and Brazilian Chamber of Deputies (PS01).

5 *Discussion and Conclusion*

Analyzing the data found in the included primary studies, it was possible to observe a growing interest in the development of research in NER for the most diverse fields of the Portuguese language, partly due to its potential applications, e.g., opinion mining, information retrieval systems, development of new general-purpose or domain-specific corpora, and optimization of machine and deep learning models. However, even with the good results achieved, the amount of research is still small when compared to other languages such as English. The amount of private or unpublished corpora and pre-trained learning models could be greater, which would improve the research area.

Looking at the research questions, it is possible to point out some limitations in the included selected studies, among which we highlight: the vast majority of studies did not deepen the discussion about the interference of the Portuguese language variant in the NER tasks; only two works showed the explicit use of a measure of agreement between manual annotators, which could influence the quality of the corpora; among the studies that used hybrid annotation, there is no comparison of the results between the types of annotation; no comparison was found between types of texts, and we think that the use of informal texts could give high complexity and richness to the textual analysis, by expressing with greater precision the colloquial form of the language; no comparison methods were found between annotation processes for entities from the same domain, it was not pointed out if there are semantic differences between entities from different domains.

Regarding the used techniques and algorithms, the most used pre-trained models

Quant.	Author	Institution	Quant.	Author	Institution
14	Renata Vieira	PUCRS-BR	5	Juliana Pirovani	UFES-BR
6	Daniela O.F. do Amaral	PUCRS-BR	4	Sandra Collovini	PUCRS-BR
6	Joaquim Santos	PUCRS-BR	3	Evandro Brasil da fonseca	PUCRS-BR
5	Bernardo Consoli	PUCRS-BR	3	Juliano Terra	PUCRS-BR
5	Elias S. Oliveira	UFES-BR	3	Nádia F. F. da Silva	UFG-BR

Table 7: Number of articles published by main researchers.

Quant.	Institution	Country	Quant.	Institution	Country
14	PUCRS	Brazil	3	IBM Research	Brazil
5	UFES	Brazil	3	PUC-Rio	Brazil
4	UFRPE	Brazil	3	UnB	Brazil
4	University of Évora	Portugal	3	UFG	Brazil

Table 8: Number of researchers per organization.

of machine learning were the classic models from the state-of-the-art, even after the advances of the recent models; more detailed information about used techniques and statistical measures was not found in the vast majority of the works.

There are some threats to the validity of our research that are worth highlighting: (i) it is possible that some relevant studies were not included throughout the search process. We attempted to mitigate this weakness by conducting extensive research as well always observing the research protocol used, carefully comparing the results and removing duplicate studies, and; (ii) as the studies were classified based on personal judgment, it is possible that some studies were classified incorrectly. In order to mitigate this threat, the classification step was performed for more than one researcher.

Finally, we emphasize that the strong point of this work is to promote the growth of research on Named Entities Recognition in the Portuguese Language, through the discrimination of their studies, resources, techniques, researchers, and institutions. We believe that the information described in this work can help other researchers/practitioners in the area to discover what has been researched and achieved, in addition to listing some gaps. The lack of some relevant data and published corpora and tools makes it difficult to carry on analysis in the research area. We plan to apply other mapping techniques to increase coverage, such as snowballing the included primary studies. Besides, an extension of this research is being produced, focusing on NER in the legislative field in several languages.

Acknowledgements

This research is carried out in the context of the Ulysses Project, of the Brazilian Chamber of Deputies. Ellen Souza and Nádia Félix are supported by FAPESP, agreement between University of São Paulo (USP) and the Brazilian Chamber of Deputies. Ricardo P. S. Filho are supported by FUNAPE. André C. P. L. F. de Carvalho and Adriano L. I. Oliveira are supported by CNPq. We express our gratitude to the Brazilian Chamber of Deputies and research funding agencies for their support this research.

References

Akbik, A., L. Chiticariu, M. Danilevsky, Y. Kbrom, Y. Li, and H. Zhu. 2016. Multilingual information extraction with polyglotie. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 268–272.

Bonifacio, L. H., P. A. Vilela, G. R. Lobato, and E. R. Fernandes. 2020. A study on the impact of intradomain finetuning of deep language models for legal named entity recognition in portuguese. In *Brazilian Conference on Intelligent Systems*, pages 648–662. Springer.

Castro, P. 2018. *Deep learning for named entity recognition in legal domain.* Ph.D. thesis, Master's thesis, Universidade Federal de Goiás.

De Araujo, P. H. L., T. E. de Campos, F. A. Braz, and N. C. da Silva. 2020. Victor:

a dataset for brazilian legal documents classification. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1449–1458.

Finatto, M. J. B., L. Lopes, and A. C. Silva. 2015. Processamento de linguagem natural, linguística de corpus e estudos linguísticos: uma parceria bem-sucedida. *Domínios de lingu@gem. Uberlândia, MG. Vol. 9, n. 5 (dez. 2015), p.[41]-59.*

Keele, S. et al. 2007. Guidelines for performing systematic literature reviews in software engineering. Technical report, Technical report, ver. 2.3 ebse technical report. ebse.

Kitchenham, B., S. Charters, et al. 2007. Guidelines for performing systematic literature reviews in software engineering version 2.3. *Engineering*, 45(4ve):1051.

Li, J., A. Sun, J. Han, and C. Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 34(1):50–70.

Maynard, D., K. Bontcheva, and I. Augenstein. 2016. Natural language processing for the semantic web. *Synthesis Lectures on the Semantic Web: Theory and Technology*, 6(2):1–194.

Nadeau, D. and S. Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

Petersen, K., R. Feldt, S. Mujtaba, and M. Mattsson. 2008. Systematic mapping studies in software engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering (EASE) 12*, pages 1–10.

Pirovani, J. P. C. 2019. *CRF+ LG: uma abordagem híbrida para o reconhecimento de entidades nomeadas em português*. Ph.D. thesis, Universidade Federal do Espírito Santo, Vitória (Brésil).

Santos, D. and N. Cardoso. 2006. A golden resource for named entity recognition in portuguese. In *International Workshop on Computational Processing of the Portuguese Language*, pages 69–79. Springer.

Souza, E., D. Costa, D. W. Castro, D. Vitório, I. Teles, R. Almeida, T. Alves, A. L. I. Oliveira, and C. Gusmão. 2018. Characterising text mining: a systematic mapping review of the portuguese language. *IET Software*, 12(2):49–75.

Souza, E., D. Vitório, D. Castro, A. L. I. Oliveira, and C. Gusmão. 2016. Characterizing opinion mining: A systematic mapping study of the portuguese language. In J. Silva, R. Ribeiro, P. Quaresma, A. Adami, and A. Branco, editors, *Computational Processing of the Portuguese Language*, pages 122–127, Cham. Springer International Publishing.

Sun, P., X. Yang, X. Zhao, and Z. Wang. 2018. An overview of named entity recognition. In *2018 International Conference on Asian Language Processing (IALP)*, pages 273–278.

Yadav, V. and S. Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470.*

A Appendix: Primary Studies (PS)

01. Albuquerque, H.O., R. Costa, G. Silvestre, E. Souza, N.F.F. Silva, D. Vitório, G. Moriyama, L. Martins, L. Soezima, A. Nunes, F. Siqueira, J.P. Tarrega, J.V. Beinotti, M. Dias, M. Silva, M. Gardini, V. Silva, A.C.P.L.F. Carvalho, and A.L.I. Oliveira. 2022. UlyssesNER-Br: A Corpus of Brazilian Legislative Documents for Named Entity Recognition. In *Proceedings of 15th edition of the International Conference on the Computational Processing of Portuguese (PROPOR 2022).* DOI: 10.1007/978-3-030-98305-5_1

02. Alves, D., B. Bekavac, and M. Tadić. 2021. The Optimization of Portuguese Named-Entity Recognition and Classification by Combining Local Grammars and Conditional Random Fields Trained with a Parsed Corpus. In *Proceedings of NooJ 2020 International Conference.* DOI: 10.1007/978-3-030-70629-6_17

03. Amaral, D., S. Collovini, A. Figueira, R. Vieira, and Marco Gonzalez. 2017. Processo de construção de um corpus anotado com Entidades Geológicas visando REN. In *Proceedings of XI Brazilian Symposium in Information and Human Language Technology and Collocated Events (STIL 2017).* Available in <https://sol.

sbc.org.br/index.php/stil/article/view/4032>.

04. Amaral, D.O.F., E.B. Fonseca, L. Lopes, and R. Vieira. 2014. Comparative Analysis of Portuguese Named Entities Recognition Tools. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).* Available in <http://www.lrec-conf.org/proceedings/lrec2014/index.html>.

05. Antonitsch, A., A. Figueira, D. Amaral, E. Fonseca, R. Vieira, and S. Collovini. 2016. Summ-it++: an enriched version of the Summ-it corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC' 16).* Available in <https://aclanthology.org/L16-1324>.

06. Batista, H.H.N., A.C.A. Nascimento, R. F. Melo, P.B.C. Miranda, I.W.S. Maldonado, and J.L.M. Coelho Filho. 2021. A comparative analysis of text embedding approach to extract named entities in Portuguese legal documents. In *Anais do Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2021).* DOI: 10.5753/eniac.2021.18255.

07. Bonifacio, L.H., P.A. Vilela, G.R. Lobato, and E.R. Fernandes. 2020. A Study on the Impact of Intradomain Finetuning of Deep Language Models for Legal Named Entity Recognition in Portuguese. In *Proceedings of 9th Brazilian Conference on Intelligent Systems (BRACIS 2020).* DOI: 10.1007/978-3-030-61377-8_46.

08. Castro, P.V.Q., N.F.F. Silva, and A.S. Soares. 2019. Contextual Representations and Semi-Supervised Named Entity Recognition for Portuguese Language. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019).* Available in <https://ceur-ws.org/Vol-2421>.

09. Collovini, S., J. Santos, B. Consoli, J. Terra, R. Vieira, P. Quaresma, M. Souza, D.B. Claro, and R. Glauber. 2019. IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 20 19).* Avaliable in <https://ceur-ws.org/Vol-2421>.

10. Consoli, B., J. Santos, D. Gomes, F. Cordeiro. R. Vieira, and V. Moreira. 2020. Embeddings for Named Entity Recognition in Geoscience Portuguese Literature. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020).* Available in <https://aclanthology.org/2020.lrec-1.568>.

11. Dias, M., J. Boné, J.C. Ferreira, R. Ribeiro, and R. Maia. 2020. Named Entity Recognition for Sensitive Data Discovery in Portuguese. Applied Sciences, 10(7):2303. DOI: 10.3390/app10072303

12. Do Amaral, D. O., E. Fonseca, L. Lopes, and R. Vieira. 2014. Comparing NERP-CRF with publicly available Portuguese named entities recognition tools. In *Proceedings of International Conference on Computational Processing of the Portuguese Language (PROPOR 2014).* DOI: 10.1007/978-3-319-09761-9_27.

13. Do Amaral, D.O.F., and R. Vieira. 2013. O Reconhecimento de Entidades Nomeadas por meio de Conditional Random Fields para a Língua Portuguesa. In *Proceedings of IX Brazilian Symposium in Information and Human Language Technology (STIL 2013).* Available in <https://sites.google.com/usp.br/stil>.

14. Do Amaral, D.O.F., M. Buffet, and R. Vieira. 2015. Comparative Analysis between Notations to Classify Named Entities using Conditional Random Fields. In *Proccedings of X Brazilian Symposium in Information and Human Language Technology and Collocated Events (STIL 2015).* Available in <https://aclanthology.org/W15-5603>.

15. Freitas, C., C. Mota, D. Santos, H.G. Oliveira, and P. Carvalho. 2010. Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10).* Avaliable in <https://aclanthology.org/L10-1284>.

16. Gonçalves, M., L. Coheur, J. Baptista, and A. Mineiro. 2020. Avaliação de Recursos Computacionais para o Porguês. *Linguamática 2020, 12, 51-68.* DOI: 10.21814/lm.12.2.331

17. Higuchi, S., C. Freitas, B. Cuconato, and A. Rademaker. 2018. Text Mining for History: first steps on building a large dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Available in <https://aclanthology.org/L18-1593>.

18. Junior, O.D., and D.B. Claro. 2011. Uma Análise do Reconhecimento Textual de Nomes de Pessoas e Organizações na Computação Forense. In *Proceddings of Sixth Internacional Conference on Forensic Computer Science (ICoFCS 2011).* DOI: 10.5769/C2011001

19. Lopes, F., C. Teixeira, H.G. Oliveira. 20 19. Named Entity Recognition in Portuguese Neurology Text Using CRF. In *Proceedings of 19th EPIA Conference on Artificial Intelligence (EPIA 2019).* DOI: 10.1007/978-3-030-30241-2_29

20. Luz de Araujo, P.H., T.E. de Campos, R.R.R. de Oliveira, M. Stauffer, S. Couto, and P. Bermejo. 2018. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. In *Proceedings of Computational Processing of the Portuguese Language (PROPOR 2018).* DOI: 10.1007/978-3-319-99722-3_32

21. Menezes, D., R. Milidiú, and P. Savarese. 2019. Building a Massive Corpus for Named Entity Recognition Using Free Open Data Sources. In *Proceedings of 8th Brazilian Conference on Intelligent Systems (BRACIS 2019).* DOI: 10.1109/BRACIS.2019.00011.

22. Miranda, N., R. Raminhos, P. Seabra, J. Sequeira, T. Gonçalves, and P. Quaresma. 2011. Named Entity Recognition using Machine Learning techniques. In *Proceedings of EPIA-11, 15th Portuguese Conference on Artificial Intelligence.* Available in <https://portulanclarin.net/static/docs/uevora-tagger/miranda2011epia.pdf>.

23. Moreira, F., and R. Vieira. 2019. Aplicação de Reconhecimento de Entidades Nomeadas em investigação de Crimes Financeiros. In *Proceedings of XII Symposium in Information and Human Language Technology and Collocates Events (STIL 2019).* Available in <http://comissoes.sbc.org.br/ce-pln/stil2019/proceedings.html>.

24. Mota, C., A. Nascimento, P. Miranda, R. Mello, I. Maldonado, and J.C. Filho. 2021. Reconhecimento de entidades nomeadas em documentos jurídicos em português utilizando redes neurais. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2021).* DOI: 10.5753/eniac.2021.18247

25. Neto, J.R.C.S.A.V.S., and T.d.P. Faleiros. 2021. Deep Active-Self Learning Applied to Named Entity Recognition. In *Proceedings of 10th Brazilian Conference on Intelligent Systems (BRACIS 2021).* DOI: 10.1007/978-3-030-91699-2_28

26. Oliveira, E., G. Dias, J. Lima, and J.P.C. Pirovani. 2021. Using Named Entities for Recognizing Family Relationships. In *Anais do IX Symposium on Knowledge Discovery, Mining and Learning.* DOI: 10.5753/kdmile.2021.17457.

27. Pinheiro, B., et al.. 2021. A Comparative Analysis of Machine Learning Named Entity Recognition Tools for the Brazilian and European Portuguese Language Variants. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional (ENIAC 2021).* DOI: 10.5753/eniac.2021.18257

28. Pires, A., J. Devezas, and S. Nunes. 2017. Benchmarking Named Entity Recognition Tools for Portuguese. *In Proceedings of Ninth INForum: Simpósio de Informática.* Available in <https://api.semanticscholar.org/CorpusID:51991813>.

29. Pirovani, J., and E. Oliveira. 2018. Portuguese Named Entity Recognition using Conditional Random Fields and Local Grammars. In *Proceedings of Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* Availabe in <https://aclanthology.org/L18-1705>.

30. Pirovani, J.P.C., and E. Oliveira. 2021. Studying the adaptation of Portuguese NER for different textual genres. *The Journal of Supercomputing, v. 77, n. 11, p. 13532-13548.* DOI: 10.1007/s11227-021-03801-9.

31. Pirovani, J.P.C., E. Oliveira. 2018. CRF+LG: A Hybrid Approach for the Portugue-

se Named Entity Recognition. In *Proceedings of 17th International Conference on Intelligent Systems Design and Applications (ISDA 2017).* DOI: 10.1007/978-3-319-76348-4_11.

32. Pirovani, J.P.C., J. Alves, M.A. Spalenza, W. Silva, C.S. Colombo, and E. Oliveira. 2019. Adapting NER (CRF+LG) for Many Textual Genres. *In Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019).* Available in <https://ceur-ws.org/Vol-2421>.

33. Quinta de Castro, P.V., N.F.F. Silva, and A.S. Soares. 2018. Portuguese Named Entity Recognition Using LSTM-CRF. In *Computational Processing of the Portuguese Language (PROPOR 2018).* DOI: 10.1007/978-3-319-99722-3_9.

34. Reyes, D.D.L., D. Trajano, I.H. Manssour, R. Vieira, and R.H. Bordini. 2021. Entity Relation Extraction from News Articles in Portuguese for Competitive Intelligence Based on BERT. *In Proceedings of 10th Brazilian Conference on Intelligent Systems (BRACIS 2021).* DOI: 10.1007/978-3-030-91699-2_31

35. Rodríguez, M.M.M.S., and B.L.D. Bezerra. 2020. Processamento de Linguagem Natural para Reconhecimento de Entidades Nomeadas em Textos Jurídicos de Atos Administrativos (Portarias). *Revista de Engenharia e Pesquisa Aplicada. 5, 1, 67-77.* DOI: 10.25286/repa.v5i1.1204.

36. Sampaio, V.A., M.J.C. França, P.B.L. Silva, G.A.L. Campos, and L.D. Hissa. 2019. A Brief Survey of Deep Learning based methods against OpenNLP NameFinder for Named Entity Recognition on Portuguese Literary Texts. In *Proceedings of XII Symposium in Information and Human Language Technology and Collocates Events (STIL 2019).* Available in <http://comissoes.sbc.org.br/ce-pln/stil2019/proceedings.html>.

37. Santos, C.N., V. Guimaraes. 2015. Boosting Named Entity Recognition with Neural Character Embeddings. *arXiv preprint.* DOI: 10.48550/arXiv.1505.05008.

38. Santos, J., B. Consoli, and R. Vieira. 20 20.Word Embedding Evaluation in Downstream Tasks and Semantic Analogies. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020).* Available in <https://aclanthology.org/2020.lrec-1.594>.

39. Santos, J., B. Consoli, C. dos Santos, J. Terra, S. Collonini, and R. Vieira. 2019. Assessing the Impact of Contextual Embeddings for Portuguese Named Entity Recognition. In *Proceedings of 8th Brazilian Conference on Intelligent Systems (BRACIS 2019).* DOI: 10.1109/BRACIS.2019.00083.

40. Santos, J., H.D.P. dos Santos, F. Tabalipa, and R. Vieira. 2021. De-Identification of Clinical Notes Using Contextualized Language Models and a Token Classifier. In *Proceedings of 10th Brazilian Conference on Intelligent Systems (BRACIS 2021).* DOI: 10.1007/978-3-030-91699-2_3.

41. Santos, J., J. Terra, B. Consoli, and R. Vieira. 2019. Multidomain Contextual Embeddings for Named Entity Recognition. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019).* Available in <https://ceur-ws.org/Vol-2421>.

42. Silva, D.F., A.M. Silva, B.M. Lopes, K.M. Johansson, F.M. Assi, J.T.C. Jesus, R.N. Mazo, D. Lucrédio, H.M. Caseli, and L. Real. 2021. Named Entity Recognition for Brazilian Portuguese Product Titles. In *Proceedings of 10th Brazilian Conference on Intelligent Systems (BRACIS 2021).* DOI: 10.1007/978-3-030-91699-2_36.

43. Silva, R.A., L. Silva, M.L. Dutra, and G.M. Araujo. 2020. A New Entity Extraction Model Based on Journalistic Brazilian Portuguese Language to Enhance Named Entity Recognition. In *Proceedings of International Conference on Data and Information in Online.* DOI: 10.1007/978-3-030-50072-6_5.

44. Souza, J.V.A., E.T.R. Schneider, J.O. Cezar, L.E.S. Oliveira, Y.B. Gumiel, E.C. Paraiso, D. Teodoro, and C.M.C.M. Barra. 2020. A Multilabel Approach to Portuguese Clinical Named Entity Recognition. *Journal of Health Informatics, Número Especial SBIS - Dezembro: 366-72.* Available in <https://jhi.sbis.org.br/index.php/jhi-sbis/article/view/840>.

45. Teteo, L., P. Moura, E. Soares, and Carlos Campos. 2019. Um Framework de Ex-

tração e Etiquetamento de Informações de Trânsito. In *Anais do XVIII Workshop em Desempenho de Sistemas Computacionais e de Comunicação*. DOI: 10.5753/wperformance.2019.6472.