Indexing Portuguese NLP Resources with PT-Pump-Up

Rúben Almeida
INESC TEC
ruben.f.almeida@inesctec.pt

Ricardo Campos
INESC TEC, Uni. Beira Interior
ricardo.campos@ubi.pt

Alípio Jorge
INESC TEC, Uni. of Porto
amjorge@fc.up.pt

Sérgio Nunes
INESC TEC, Uni. of Porto
ssn@fe.up.pt

Abstract

The recent advances in natural language processing (NLP) are linked to training processes that require vast amounts of corpora. Access to this data is commonly not a trivial process due to resource dispersion and the need to maintain these infrastructures online and up-to-date. New developments in NLP are often compromised due to the scarcity of data or lack of a shared repository that works as an entry point to the community. This is especially true in low and mid-resource languages, such as Portuguese, which lack data and proper resource management infrastructures. In this work, we propose PT-Pump-Up, a set of tools that aim to reduce resource dispersion and improve the accessibility to Portuguese NLP resources. Our proposal is divided into four software components: a) a web platform to list the available resources; b) a client-side Python package to simplify the loading of Portuguese NLP resources; c) an administrative Python package to manage the platform and d) a public GitHub repository to foster future collaboration and contributions. All four components are accessible using: `https://linktr.ee/pt_pump_up`

1 Introduction

The topic of NLP resource management in European languages was initially introduced by Danzin (1992), with the first references to Portuguese resources presented ten years later in the works of Santos (2002). The recent advances in NLP, linked to the development of large language models reintroduced the debate about NLP resource management due to the large volume of training data required by these architectures. Several platforms have been recenlty introduced offering different approaches to address this problem. In Table 1, we present some of these platforms focusing on those that index, to some extent, Portuguese NLP resources. Our analysis reveals that Portuguese NLP resources are dispersed across more than 11 platforms which implement different storage policies; some only store metadata, while others focus on providing entire copies of the resources indexed. In a mid-resource language such as Portuguese (Joshi et al., 2020), this resource dispersion phenomenon exacerbates the already existing challenges linked to the reduced amount of NLP resources, negatively impacting the accessibility to these resources.

To address these challenges, we extend the surveying works of Almeida (2023) and propose PT-Pump-Up, a set of tools that support the development of the first centralising platform for Portuguese NLP resources. In this demonstration, we present the minimum set of valuable features to achieve this goal divided across the four software components that compose PT-Pump-Up: a) a web platform; b) a client Python package; c) an administrative Python package and d) a public GitHub repository. Additional details about this release are available in the wiki of the project.

2 PT-Pump-Up

The architecture of PT-Pump-Up is represented in Figure 1, which highlights not only the features already implemented but also the work in progress and future plans associated with this project. In this demonstration, we present four scenarios where PT-Pump-Up can be employed to mitigate resource dispersion and enhance synchronization across diverse platforms that support Portuguese NLP resources.

2.1 Easy Access to Portuguese NLP Resources

The PT-Pump-Up Python client[1] permits the easy loading of Portuguese NLP resources. The resource is loaded directly if it has a copy in HuggingFace[2]; if not, it returns the metadata that describes it. In Listing 1, we demo how to use PT-Pump-Up to achieve this goal using a few lines of code.

[1] `https://pypi.org/project/pt-pump-up`
[2] `https://huggingface.co`

Platform	Updated	Origin	# PT Var.	Colab.	Meta.	Res.
Portulan Clarin (Branco et al., 2023)	✓	PT	1	⚠	⚠	✓
NILC: Tools and Resources	✓	BR	1	×	×	✓
Portuguese-NLP	✓	BR	⚠	✓	✓	×
HuggingFace	✓	FR	⚠	✓	⚠	✓
PapersWithCode	✓	USA	⚠	✓	✓	×
ELRA	✓	BE	1	×	×	✓
Open Language Archives Community (Simons et al., 2003)	✓	USA	2+	×	✓	×
AiLab	2021	BR	⚠	✓	✓	×
ACL Wiki: Resources for Portuguese	2020	USA	⚠	×	✓	×
Organização Etica.AI	2018	BR	⚠	×	✓	×
Linguateca (Santos et al., 2004)	2012	PT	⚠	×	×	✓
PT-Pump-Up	✓	PT	2+	✓	✓	⚠

Table 1: Overview of NLP resource management platforms supporting Portuguese resources, emphasizing discontinued projects (shaded in grey). Details encompass platform origin, collaborative operation (Colab.), and storage of resource metadata (Meta.) or complete copies of resources (Res.). The warning sign indicates omitted or incomplete information. For the count of Portuguese Varieties indexed (#PT Var), "2+" denotes platforms considering at least the two predominant Portuguese varieties: European and Brazilian.

```python
from pt_pump_up.PT_Pump_Up import PTPumpUp
client = PTPumpUp()
all_ner_datasets =
 ↪  client.all_datasets(nlp_task="Named Entity
 ↪  Recognition")
print(all_ner_datasets.head())
# Dataset is Loaded as a HF Dataset object
dataset = client.load_dataset(english_name=str)
```

Listing 1: Load Portuguese named entity recognition dataset.

2.2 Indexing Portuguese NLP Resource

The PT-Pump-Up administrative package permits authenticated CRUD operations to manage the resources indexed in the platform. These actions can also be done using the web interface, ensuring that the absence of programming skills is not a barrier to interacting with the platform. In Listing 2, we demonstrate how a new NLP task can be included in PT-Pump-Up with a few lines of Python code.

2.3 Measure Resource Preservation Needs

We introduced a *resource preservation rating* to identify less accessible resources. This metric enabled the establishment of a hybrid storage policy. For resources with high preservation ratings, only metadata is stored, while those with lower ratings are prioritized for human intervention and the creation of a backup copy. The preservation rating can be provided during resource submission or automatically determined using a decision tree integrated into the admin package.

```python
from pt_pump_up_admin.PT_Pump_Up import
 ↪  PT_Pump_Up
from pt_pump_up_admin.crud.NLPTask import
 ↪  NLPTask

pt_pump_up = PT_Pump_Up(
    # Get token: pt-pump-up.inesctec.pt/dashboard
    bearer_token=str,
)
NLPTask(pt_pump_up=pt_pump_up).insert(
    name=str,
    acronym=str,
    papers_with_code_ids=list,
)
```

Listing 2: Inserting a NLP task to the database.

```python
#TODO: Mudar estes imports para o Admin package
from pt_pump_up.rating.Preservation_Rating
 ↪  import PreservationRatingDataset

# Instantiates a client
client = PreservationRatingDataset()

preservation_rating = client.predict(#...dataset
 ↪  proprieties)
print(preservation_rating)
```

Listing 3: Predicting preservation rating of a dataset based on its metadata

2.4 Integrate With Papers With Code

The PT-Pump-Up integration module included in the admin package compresses the logic developed to enforce resource synchronization with other platforms. In this release, we deliver the tools to sup-

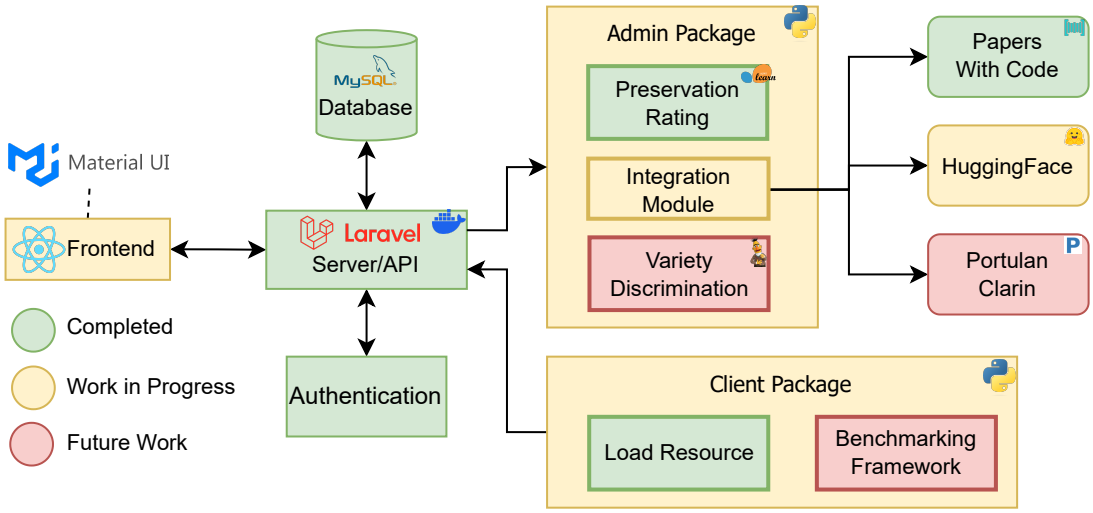

Figure 1: Architecture of PT-Pump-Up. Background colours highlight the completeness of each module.

port the integration with Papers With Code. This module presents challenges due to the heterogeneity of systems used by the targeted applications. In Listing 4, we demonstrate how PT-Pump-Up can be used to synchronise a resource with Papers With Code using a few lines of code.

```
from pt_pump_up.papers_with_code.PapersWithCode
↪  import PapersWithCodeDataset, PapersWithCode

# Login in PapersWithCode
client = PapersWithCode(username=str,
↪  password=str)

#Create Dataset instance
dataset = PapersWithCodeDataset(#...dataset
↪  proprieties)
#Publish Resource
client.insert(dataset)
```

Listing 4: Insert dataset metadata to Papers With Code.

3 Conclusion and Future Work

This paper details the first release of PT-Pump-Up and how its tools can be used to address the challenge of Portuguese NLP resource dispersion. In this release, we deliver the minimum valuable set of essential features capable of demonstrating the four software modules that compose PT-Pump-Up. This project is a work in progress, with future work focusing on extending the integration module to other platforms and extending the resources indexed with the collaboration of the Portuguese NLP community.

Acknowledgements

This work is financed by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. within the project StorySense, reference 2022.09312.PTDC(DOI 10.54499/2022.09312.PTDC).

References

Rúben Almeida. 2023. Building portuguese language resources for natural language processing tasks. MSc Thesis, Faculty of Engineering, University of Porto.

António Branco et al. 2023. The clarin infrastructure as an interoperable language technology platform for ssh and beyond. *Language Resources and Evaluation*, pages 1–32.

A Danzin. 1992. Towards a european language infrastructure (dg xiii).

Pratik Joshi et al. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. *CoRR*, abs/2004.09095.

Diana Santos. 2002. Um centro de recursos para o processamento computacional do português. *DataGramaZero-Revista de Ciência da informaçao*, 3(1).

Diana Santos et al. 2004. Linguateca: um centro de recursos distribuído para o processamento computacional da língua portuguesa.

Gary Simons et al. 2003. The open language archives community: An infrastructure for distributed archiving of language resources. *Literary and Linguistic Computing*, 18(2):117–128.