FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Building Portuguese Language Resources for Natural Language Processing Tasks

Rúben Almeida

Mestrado em Engenharia Informática e Computação

Supervisor: Alípio Jorge

Second Supervisor: Sérgio Nunes

July 26, 2023

Building Portuguese Language Resources for Natural Language Processing Tasks

Rúben Almeida

Mestrado em Engenharia Informática e Computação

July 26, 2023

Abstract

Natural Language Processing (NLP) is a rapidly advancing field in Computer Science, mainly due to the significant developments in Deep Learning. These advances are, unfortunately, asymmetric across the different human languages. The access to large volumes of annotated data in resource-intensive languages like English and Mandarin simplified the training of state-of-art models that require vast sums of corpora. In contrast, low-resource languages, like Portuguese, have less labelling schemes (LS), negatively impacting the existing models' quantity and quality.

Despite these challenges, there are Portuguese NLP resources available (datasets and models), but they are dispersed among various research teams, countries, variants, platforms and follow different annotation schemes. To address this dispersion, we conducted a comprehensive literature review to catalogue them and determine the current benchmarks for a subset of 8 NLP tasks. Our findings confirmed our motivations, the resources are scarce and dispersed, do not follow the most recent LS, and most solutions do not prioritise an off-the-shelf approach.

To overcome these challenges and contribute to normalising the asymmetries for the Portuguese case, we exploited paths to improve the process of fine-tuning pre-trained transformer-based architectures by creating larger corpora based on the combination of smaller corpus with similar LS. However, there are several Portuguese NLP tasks without any corpus. In those cases, we propose using Data Augmentation (DA) based on automatic machine translation to generate silver-labelled data.

We validated this approach by researching Named Entity Recognition (NER) and Abstractive Text Summarization (ATS). First, we developed a training pipeline for NER that fine-tunes a BERT-CRF architecture in three LS: HAREM Selective, HAREM Default and CoNLL-2003. Our work achieved a new benchmark of 0.951 F1-Score on the Wikineural dataset. Combining datasets revealed insufficient to outperform current results in the other two LS; this motivated us to use our DA pipeline to translate the Ontonotes 5.0 dataset, which shares a similar LS with HAREM. The results were negatively impacted by the token alignment solutions used.

The conclusions drawn from our NER research prompted adaptations in our approach to ATS. Since this task does not require a token alignment step, we performed further research on the impact of silver-labelled data in the training process. We translated a sample of twenty-five thousand documents from the English dataset CNN-Dailymail and compared the ROUGE-L scores obtained under the same training conditions for original and translated data. The results were identical, motivating us to continue our research in DA as a relevant form of extending Portuguese NLP resources.

Finally, we included automatic deployment features in the pipelines developed to ensure that all the models and datasets produced during this dissertation are available on Huggingface and in GitHub in an off-the-shelf manner.

Keywords: Portuguese NLP, Named Entity Recognition, Text Summarization, Data Augmentation, Low-Resource Languages

Resumo

O Processamento de Linguagem Natural (PLN) é uma área em crescimento na Ciência da Computação, impulsionada pelo avanço das redes neuronais profundas. Porém, a disponibilidade e qualidade dos recursos de PLN em Português são limitadas em comparação com o Inglês e o Mandarim.

Apesar destes desafios, existem recursos (corpus e modelos) para PLN Português, estando estes dispersos por várias equipas de investigação, países, variações do Português, plataformas e esquemas de anotação (LS). Para contrariar esta dispersão fizemos uma extensa revisão de literatura visando catalogar estes recursos para uma amostra de 8 tarefas de PLN. Os resultados obtidos confirmam as nossas premissas, os recursos são, de facto, escassos e estão dispersos, não seguem os LS mais recentes e tendem a não priorizar serem soluções *off-the-shelf*. Para reduzir estas adversidades, introduzimos uma metodologia para melhorar o processo de *fine-tuning* de modelos pretreinados baseados em *transformers* através da combinação de corpus de menor dimensão numa unidade de treino maior. Além disso, exploramos a técnica de Data Augmentation (DA) com base em tradução automática sempre que se verifica que há escassez de dados para uma tarefa específica.

Validamos esta abordagem, em Reconhecimento de Entidades Nomeadas (NER) e Sumarização Abstrativa (ATS). Primeiro, desenvolvemos uma *pipeline* de treino para NER baseada no *fine-tuning* da arquitetura BERT-CRF em três LS– HAREM Seletivo, HAREM *Default* e CoNLL-2003. O nosso trabalho permitiu atingir um novo *benchmark* de 0.951 F1-Score usando o corpus Wikineural. A metodologia de combinação de corpus revelou-se insuficiente para ultrapassar os melhores resultados existentes para os restantes LS. Este facto motivou-nos a utilizar a *pipeline* de DA desenvolvida para traduzir o corpus Inglês Ontonotes 5.0. Infelizmente, os resultados obtidos foram impactados negativamente pela fraca qualidade dos modelos disponíveis para alinhamento de *tokens*.

As conclusões retiradas do processo de desenvolvimento para NER, motivou alterações na forma como abordamos ATS. Aproveitamos o facto desta tarefa de PLN não necessitar de alinhamentos de *tokens* para explorarmos mais a fundo a inclusão de dados sintéticos no processo de treino. Para isso, traduzimos uma amostra do dataset CNN-Dailymail e comparamos os resultados de ROUGE-L obtidos nas mesmas condições em dados originais e sintéticos. Os resultados foram semelhantes, o que nos motiva a continuar a nossa investigação de como utilizar DA para contrariar a escassez de recursos de PLN em língua portuguesa.

Finalmente, publicamos todos as pipelines, modelos e corpus produzidos durante esta dissertação no HuggingFace e no GitHub garantindo que estas soluções estão prontas para ser utilizadas de uma forma acessível e *off-the-shelf*.

Palavras-Chave: Processamento de Linguagem Natural, Reconhecimento de Entidades Nomeadas, Sumarização, Dados Sintéticos, Linguagens de Baixos Recursos

Agradecimentos

Firstly, I must thank my supervisors, Prof. Alípio Jorge and Prof. Sérgio Nunes, for the opportunity and assistance provided during all these months of research.

Then, I would like to thank all the elements of the NLP research group at the Faculty of Science of the University of Porto that each Thursday provided me good insights on how to produce a quality dissertation.

Não posso deixar de esquecer a minha família e quero, desde já, agradecer-lhes pelo almoço que tem de me pagar depois da apresentação pública. Quero deixar uma memória para a pessoa que mais suporte deu à família nos momentos mais complicados, o meu falecido avô Celestino.

Last but not least, I consider it relevant to thank the two most important peers during my five years at university. Thank you, Manuel Coutinho and José Guerra. I wish you all the best in your lives.

Muito obrigado a todos :)

Rúben Filipe Seabra de Almeida

"A Europa jaz, posta nos cotovelos:
De Oriente a Ocidente jaz, fitando...
O rosto com que fita é Portugal."

Fernando Pessoa, in Mensagem. O dos Castelos

Contents

List of Figures

List of Tables

Chapter 1

Introduction

We begin by introducing the context (1.1) and the research project this dissertation is part of, the PT-Pump-Up (1.2). Then we state the problem that motivates it (1.3) and the goals we intend to achieve (1.4). Lastly, we explain the structure of this document (1.5).

1.1 Context

Natural Language Processing (NLP) is the Computer Science (CS) field concerned with turning computers capable of processing and understanding human language [46]. It is one of the topics that shows a more significant increase in economic value [1] and the amount of research produced in recent years. With an annual economic growth of 20% and a valuation of $161.8 in 2029 [1], the capabilities of automatic information extraction from corpora are changing CS.

NLP was born in the 1940s in the United States with the efforts of creating machine translation systems [96]. Quickly scientists faced the challenges human languages impose due to their embedded ambiguity and diversity [12]. Over the year, the field has observed gradual improvements. However, the latest paradigm based on transformer models [125] beat several benchmarks, accelerating the growth momentum that NLP was already observing.

NLP is a dynamic research field that includes several tasks related to human language processing. The website Papers With Code identifies 578 different NLP tasks[1]. The list provided by Papers With code is extensive and in continuous expansion. It includes not only NLP tasks that are widely used in modern societies like Machine Translation (2.2.9), but also many other tasks less known by end-users, yet relevant, like Temporal Information Extraction (2.2.6)

The increased demand for NLP solutions to solve CS problems outside the traditional NLP scope introduced several libraries that cover different NLP tasks in multiple languages. These tools deliver high levels of abstraction to tackle the specificity of NLP development that broadens

[1] https://paperswithcode.com/area/natural-language-processing

the span of users capable of interacting with them. Tools like spaCy[2] or HuggingFace[3] offer several off-the-shelf solutions, which justifies millions of downloads each month[4].

In this document, we focus our attention on the Portuguese language. This language has a significant number of variants. Nevertheless, most NLP resources are introduced in only two variants, the European and the Brazilian cases. This dissertation focuses on European Portuguese, yet we ensure that the Portuguese variant is explicitly identified whenever it is essential. If only Portuguese is mentioned, we refer to both the European and the Brazilian cases.

The research in Portuguese NLP introduces several challenges, and our proposals to solve them establishes the foundations for the beginning of the PT-Pump-Up project (1.2). We propose two lines of research. The first is the cataloguing of Portuguese resources. The second is the extension of the resources available in Named Entity Recognition and Text Summarization.

1.2 PT-Pump-Up Project

The PT-Pump-Up project and initiative aim to boost NLP research for Portuguese. By introducing a public repository that catalogs NLP resources and developing methods to support the automatic adaptation of resources from other languages, PT-Pump-Up is expected to become a primary source of reliable NLP resources in Portuguese. PT-Pump-Up starts with the developments made during this master thesis. The contributions made to it, and the project are interchangeable since they contribute to the success of both assignments.

1.3 Motivation

In the following subsections, we list the four significant problems that motivate this dissertation.

1.3.1 Inequality Among Different Languages

The progress in NLP is different for each human language. We considered the main reason for these differences to be the number of language speakers. There are 200 million Portuguese speakers worldwide [2], a small value compared to languages like English or Mandarin. As a consequence of this figure, we identify two negative impacts on NLP Portuguese research. The first is that speakers produce fewer written resources. The reduced number of web pages, social network posts, and newspapers in Portuguese negatively impacts the accessibility to Portuguese corpora for researchers to produce work. The second consequence is less investment in producing resources in Portuguese NLP. Compared to English and Mandarin, the smaller market moves away major IT companies and, with them, their significant amount of financial, data and talent. The consequence of these facts is fewer NLP resources for the Portuguese language. Not only are they less, but they

[2] https://spacy.io/
[3] https://huggingface.co/
[4] https://pypistats.org/packages/spacy

are typically dispersed across platforms and research teams, making the research in Portuguese NLP more challenging and time-consuming than in other languages.

A quick search on HuggingFace[5], a popular platform in the field, introduces figures that illustrate this problem. From the 16,676 datasets available on this platform, 2,044 are in English, and only 179 are marked as Portuguese. Of these 179, many of them are multi-language. These figures clearly show this inequality. The same problem transposes to models; from the 106,746 models available for download on this platform, 10,762 are trained for the English Language. In the Portuguese case, only 541 models are available. Figures below (1.1, 1.2, 1.3) present a comparison between the resources available on major online platforms and the number of worldwide speakers in each language.

Figure 1.1: Relation between number of speakers and HuggingFace datasets. Excluding English

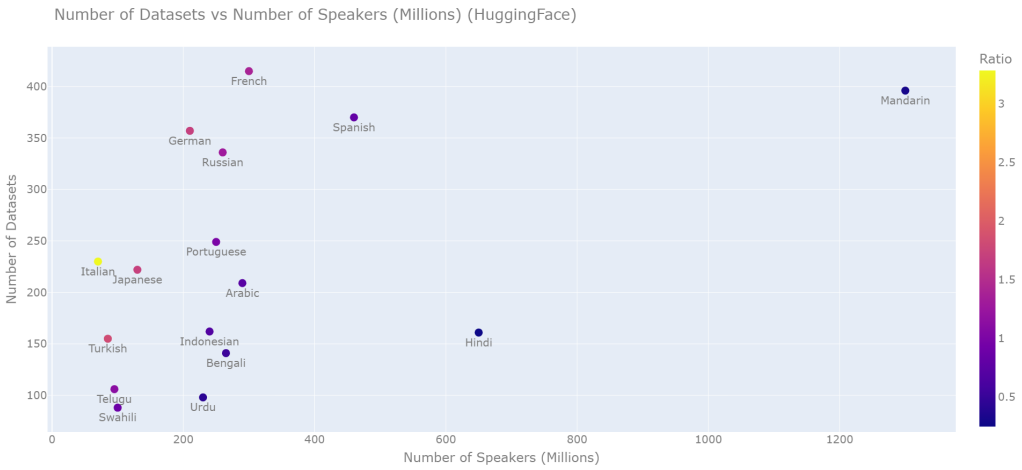
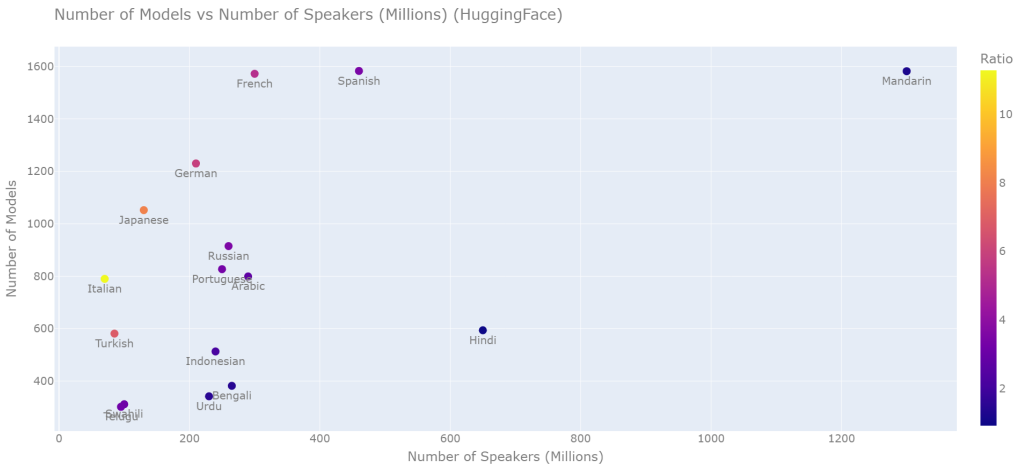
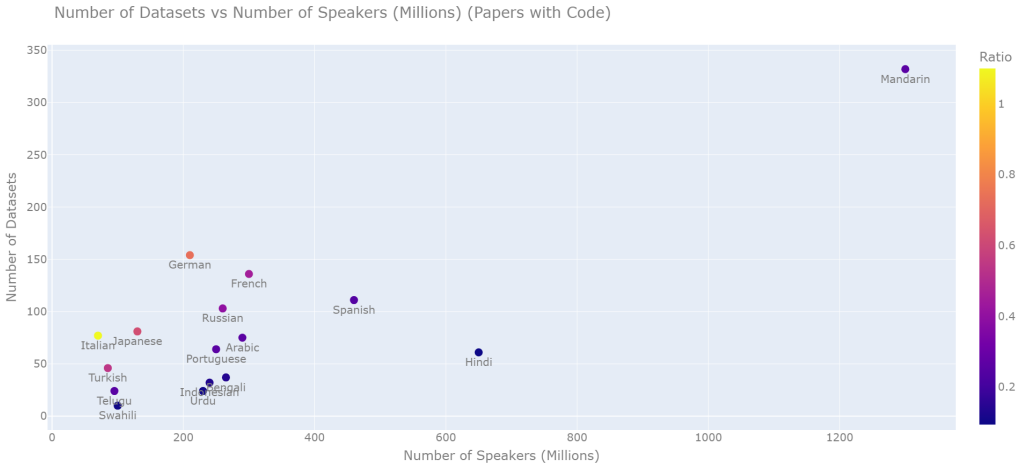


Figure 1.2: Relation between number of speakers and HuggingFace models. Excluding English



[5] https://huggingface.co/

Figure 1.3: Relation between number of speakers and Papers With Code datasets. Excluding English



The three plots (1.1, 1.2, 1.3) do not clearly show a recognisable pattern relating the number of speakers and the number of resources available on this platform, showing, instead, the incredible amount of work still required outside the English language to equalise these figures.

1.3.2 Challenges for Transformers Solutions in the Portuguese Language

The results obtained by transformer models empowered a line of thought in NLP that the bigger the model, the better the results [125]. A bigger model requires more extensive datasets, more computation resources and training time. All these three aspects constitute an obstacle to Portuguese NLP research:

Lack of Datasets State-of-the-art models like GPT-3 are trained with terabytes of information. The resources in Portuguese are scarce and of small size, which is incompatible with the requirements of transformer architectures. This limitation makes training these architectures in the Portuguese case almost impossible.

Lack of Computing Resources Transformers require extensive computing resources for their training. Millions of dollars are required in the training process. Major IT companies typically sponsor the training of the most powerful architectures. The combination of fewer major tech companies in Portuguese-speaking countries, with a bigger economic motivation to produce for bigger markets like the United States and China, pushes the companies based in Portuguese-speaking countries away from developing modern solutions specialised in the Portuguese case.

1.3.3 Cross-Language Models

To cover several languages, well-known NLP libraries like spaCy[6] and Stanza[7] train multi-language models. Nevertheless, these solutions focus on languages with more significant markets, like the English-speaking world, rather than the Portuguese-speaking case. Most of the time, the Portuguese solutions those libraries provide lack completeness and documentation and are trained in poor-quality and outdated datasets. The immediate consequence of these facts is that the Portuguese off-the-shelf model performs lower benchmarks than solutions in resource-intensive languages.

1.3.4 Portuguese Datasets Require Standard Annotation Formats

Most of the 578 NLP tasks catalogued by Papers With Code[8] have specific datasets used as a reference for benchmarking. For example, for question answering, the SOTA reference dataset for testing is SQuAD2.0 [93], and for sentiment analysis, Amazon Reviews is widely used [42]. The format these datasets adopt tends to create unofficial standards. The CoNLL-2003 format is a good example of an unofficial annotation convention for Token Classification tasks like Part-of-Speech Tagging (POS) and Named Entity Recognition (NER) datasets.

The lack of standardisation in the Portuguese case affects the task of cataloguing made by major platforms like Papers With Code[9] and HuggingFace[10] since these tools make an effort to present only well-formatted and standardised datasets. Not having these resources available on major platforms impact its easiness of access negatively. In order to simplify the publishing process of the existent Portuguese datasets, a normalisation process is required. This normalisation process is not trivial, and it is time-consuming. Therefore, many of the Portuguese datasets remain to be normalised and catalogued by these platforms.

A circular reasoning problem arises since organisations and individuals neglect the development of Portuguese NLP solutions due to its higher overhead in accessing resources, pushing them away from research and compromising the introduction of new resources that the community lacks.

1.4 Goals

Despite the problems listed above, there is plenty of room to contribute for developing NLP research in Portuguese language. The simple transposition of modern architectures produced for languages like English to Portuguese would already contribute significantly.

The recent advances in NLP, namely transformers [125], opened a broad range of possibilities yet to be explored. There is literature [67, 91] that shows how transformers can address the

[6] https://spacy.io/
[7] https://stanfordnlp.github.io/stanza/
[8] https://paperswithcode.com/
[9] https://paperswithcode.com/
[10] https://huggingface.co/

problem of lack of datasets in low-resource languages. New Data Augmentation techniques, using transformer text generation capabilities to create more complete datasets, resilient to variance and bias, show promising results [91]. Exploiting the adaptability of transformer architecture to create models more adapted to a specific context using fine tuning techniques also shows promising results for low resources languages [91].
We identify four primary goals for PT-Pump-Up, based on the problems previously identified (1.3).

1.4.1 Develop a Public Repository for Portuguese NLP

To address the need for resource cataloguing in the Portuguese NLP community, it is expected that the extensive information gathering during our literature review (3) will be the object of a new public access repository. In addition to the resources identified, all the contributions made during this project to extend the existing resources will be published on this platform. This new Portuguese NLP hub is expected to include automation to support publishing in major platforms like Papers With Code[11] or HuggingFace[12]. Despite the importance of Portulan Clarin[13], this platform will be dealt with manually, given its current way of operation.

1.4.2 Datasets Extension

As mentioned, modern transformer architecture produces exciting results based on large amounts of data, but the few pre-existent Portuguese datasets are small. New techniques based on Data Augmentation can mitigate this problem. We intend to extend the number and the quality of the data resources available for the Portuguese case by taking advantage of these techniques. We intend to use Machine Translation (2.2.9) models to use resources in other languages to augment the number of Portuguese corpora available for training.

The intuition is that bigger and better data sources facilitate the achievement of higher model training results. All relevant contributions in this field will be published publicly in the PT-Pump-Up platform to support the research of the Portuguese NLP community.

1.4.3 Producing Off-the-Shelf Models

One of the problems we intend to help solve is the reduced number of off-the-shelf models ready to use, fine-tune or adapt for most Portuguese NLP tasks. We consider that valuable NLP solutions are also the ones that are accessible to use; therefore, we will enforce that all the models trained for NER and ATS will be deployed in major NLP frameworks, simplifying their usage. We intend to deploy our contributions in HuggingFace[14].

[11] https://paperswithcode.com/
[12] https://huggingface.co/
[13] https://portulanclarin.net/
[14] https://huggingface.co/

1.4.4 Elevate SOTA Benchmarks of Portuguese NLP Tasks

To address the lower benchmarks offered by off-the-shelf Portuguese NLP tools, we expect to elevate the benchmarks for Named Entity Recognition (5) and Abstractive Text Summarization (6) by applying modern neural architectures combined with silver-labelled data derived from the techniques mentioned in the previous Subsection (1.4.2).

1.4.5 Summary

The solutions proposed have different levels of technical complexity. The process of surveying Portuguese NLP resources offers little technical complexity, yet it is time-consuming. Others, like the lack of transposition of SOTA techniques to the Portuguese language, require an extensive understanding of the literature that supports those architectures. Nevertheless, we believe that all contributions described in this document (7.3), independently of their technical requirements, are relevant to the Portuguese NLP community.

1.5 Document Structure

This document is composed of seven chapters, structured as follows:

Chapter 1. Introduction Introduces the problem we propose to address and describes the context and the motivations that define this dissertation.

Chapter 2. Theoretical Background Covers the core concepts of NLP we consider relevant to better understand the methodology followed and the results obtained.

Chapter 3. Portuguese Natural Language Processing Includes the literature review of the different Portuguese resources in eight NLP tasks.

Chapter 4. Building Portuguese Natural Language Resources Explains the methodology proposed to achieve the planned goals. Formulates the research questions, the main hypothesis and clarifies some architectural decisions.

Chapter 5. Portuguese Named Entity Recognition Presents the results obtained by the NER pipelines developed.

Chapter 6. Portuguese Abstractive Text Summarization Presents the results obtained by the ATS pipelines developed.

Chapter 7. Conclusions Concludes this document by presenting the significant takeaways and the identification of future work topics.

Chapter 2

Theoretical Background

In the previous chapter, we briefly introduced the problem we propose to address, the goals we intend to achieve and the motivations to do so. This chapter provides a theoretical contextualisation of the concepts presented in this dissertation. Readers can skip this section if they consider their knowledge of the concepts presented solid.

We start by providing an introduction to the general concepts of NLP. We contextualise modern NLP as a Machine Learning problem (2.1.1). Then, we present a classification for NLP tasks (2.1.2), and then we clarify the difference between low and high-level NLP tasks (2.1.3). This section concludes by introducing sequence-to-sequence tasks, a concept deeply associated with NLP (2.1.4).

In Section 2.2, we briefly describe the eight NLP tasks we propose to catalogue the existing resources in Portuguese.

2.1 Natural Language Processing

In this section, we focus our attention on essential topics in NLP. First, we abstract modern NLP as an ML problem by providing a taxonomy to clarify the classification of the many NLP research fields (2.2). After that, we clarify the concepts of low and high-level NLP tasks (2.1.3) and describe what a sequence-to-sequence task is (2.1.4).

2.1.1 Machine Learning Algorithms Trending in NLP

Like most AI fields, natural language processing is currently dominated by ML techniques due to the quality of the results presented by these models. The application of ML in NLP introduced neural networks, word embeddings and, more recently, attention-based transformers models. All these methods constitute the basis of modern NLP. Figure 2.1 presents a chronology of SOTA techniques used in NLP.

Figure 2.1 evidences an increase in research developed in NLP after the introduction of Transformer models. The results obtained by deep-learning techniques appear to have motivated the community to introduce new contributions.

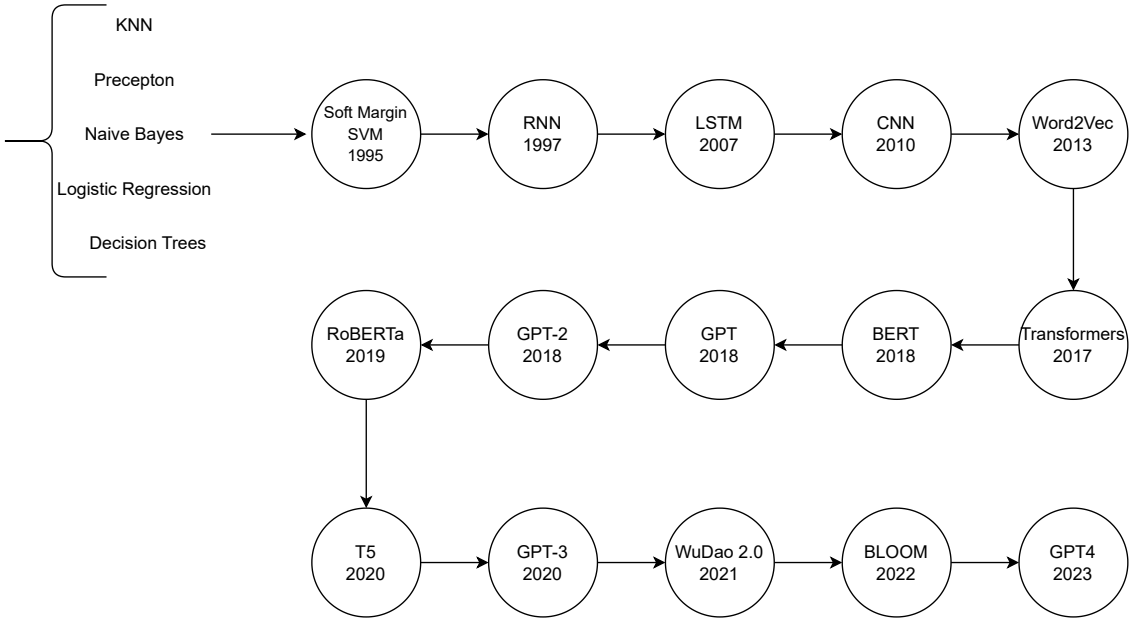


Figure 2.1: Temporal analysis of ML algorithms trending in NLP.

2.1.2 Different Types of NLP Tasks

Natural Language Processing is a research field that includes several different tasks. Figure 2.2 present a schematic that aims to provide a high-level, non-extensive visual analysis of the diversity of NLP tasks. Our classification highlights six major categories of NLP tasks:

Knowledge Graphs Related Tasks A Knowledge Graph is a graph-based representation of entities and the relation between them. Knowledge Graph tasks aggregates research related to the extraction and linking of entities [44].

Text Classification Text Classification tasks are a field of NLP primarily associated with predicting a correct label for an input prompt. Tasks like sentiment analysis, spam and fake-news detection are good examples of these tasks [50].

Text-to-Data and Data-to-Text Text-to-Data and Data-to-Text includes the NLP research that intends to extend and integrate NLP with non-conventional textual inputs. Tasks like Speech-to-Text are excellent examples of the integration of NLP with the world of phonetics and electrical signalling processing [18].

Text Generation Text Generation is one of the most relevant areas of research in NLP. Some taxonomies categorise this field as Natural Language Generation (NLG) [94]. It includes the capabilities of producing coherent artificial text.

Text Pre-Processing Text Pre-Processing is the most used field of NLP. It includes tasks that serve as the basis for more complex NLP challenges. The process of preparing, modelling and filtering corpora is present in every NLP pipeline [51].

Information Retrieval Information Retrieval(IR) is a field of research independent of NLP. It includes tasks associated with querying and document retrieving from extensive collections. However, the recent advances in NLP forced IR to embrace several NLP techniques to improve the results obtained in this field [19].

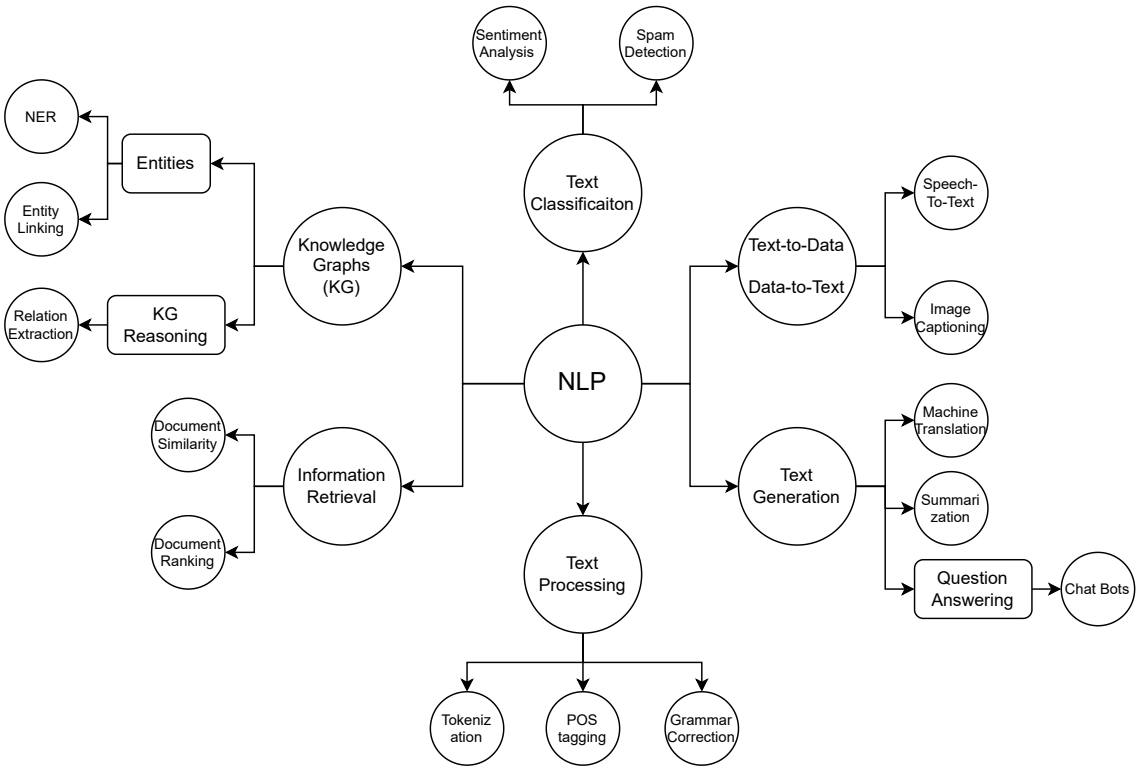


Figure 2.2: NLP tasks taxonomy [103].

Nature Language Processing tasks are typically complex problems that span across different fields. For example, Question Answering includes several entity processing features and text generation capabilities. The taxonomy proposed is an element of awareness of the diversity of NLP tasks rather than the best organisation for taxonomise NLP tasks.

2.1.3 Difference Between Low and High-Level NLP Tasks

In this document, we follow a naming convention to better understand the ultimate goal of a specific NLP task. We name *Low-Level NLP Task*, a task that performs a specific purpose deeply associated with the syntactic construction of phrases. In contrast, a *High-Level Task* typically uses several low-level tasks to perform more general-purpose tasks with a higher level of complexity.

Examples of low levels tasks are:

- Part of Speech Tagging
- Semantic Role Labelling
- Temporal Expression Extraction

Examples of high-level tasks are:

- Text Summarization
- Question Answering
- Machine Translation

2.1.4 Sequence To Sequence Tasks

As already mentioned, NLP is currently an ML problem due to the improved performance of SOTA statistical ML models. Nevertheless, NLP introduces specificities to a generic ML problem. One of those specificities is the introduction of tasks of sequence labelling [13]. Sequence Labeling tasks "assign a label chosen from a small fixed set of labels to each sequence element" [13]. Due to the sequential nature of human languages, it is easy to frame tasks like NER, POS and Machine Translation as sequence labelling problems.

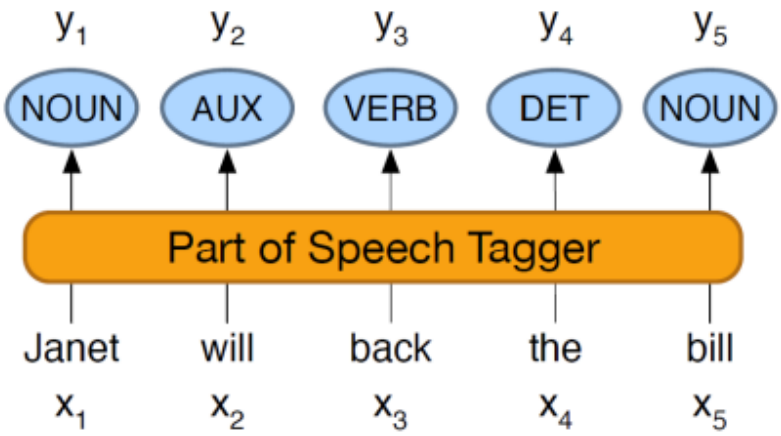


Figure 2.3: Part of speech tagging as a sequence labelling task [13].

When an NLP task can be abstracted as a sequence labelling problem, models can extract additional information from the input text. The configuration of words in well-formed human languages sentences is typically related. For example, in German, the verb is always the second POS of a sentence, and the subject must be placed next to the verb. By taking advantage of the sequential nature of this problem, German language POS models can extract relevant information from the position of the tokens in a sentence.

2.2 Natural Language Processing Tasks

In the previous sections of this document, we contextualised modern NLP as an ML problem (2.1.1). In this section, we focus our attention on briefly introducing the tasks that will be covered in the state-of-the-art Section of this document (3) with particular attention to Text Summarization (2.2.2) and Named Entity Recognition (NER) (2.2.4) since those are the object of the implementation Sections of this document (5, 6).

Moreover, depending on the research team, variants exist in the naming given to a specific NLP task. In order to remove this type of ambiguity, we decided to follow the naming of the website Papers With Code[1].

2.2.1 Data Augmentation

Data augmentation (DA) "encompasses methods of increasing training data diversity without directly collecting more data" [35]. DA formally augments data by applying one or more transformations to a gold-labelled dataset. In those processes, the labels must be guaranteed to preserve their initial meaning and order [127]. The wide range of valid transformations motivated several recent advances in DA techniques.

The most classical DA technique is back-translation [107], which uses automated machine translation systems (2.2.9) to produce synthetically generated data. Synonyms replacement [129] is also a widely used technique due to its low computational overhead. Data Augmentation techniques observe an "increase interest and demand" [35] due to the inequality in the amount and quality of corpora available for low-resource languages.

Unfortunately, applying DA techniques to some NLP tasks is not straightforward. In particular, for seq-2-seq tasks like NER, it is necessary to ensure that the DA technique does not modify the length and order of the tokens in a sentence. This limitation restrains the range of DA techniques available. For example, DA based on Machine Translation turns more challenging since the length, and the order of the tokens are not always the same across different languages (Ex: "Happy man" in Portuguese is spelt in reversed order "Homem feliz"). To overcome this situation, several literature was introduced in recent years; we explore this subject in more detail in Section 3.2.

In recent years, other non-conventional DA techniques have been developed, and the most promising ones use the recent advances in transformers models. New DA techniques exploit the Text Generation capability of LLMs to generate valuable synthetic data.

2.2.2 Text Summarization

Text Summarization is the NLP task capable of creating concise summaries of one or more text documents [60]. H. Lin and V. Ng present Automatic Text Summarization as a "significant contribution to modern society due to the information overflow phenomena" [60]. There are two significant types of summarization, abstractive and extractive summarization.

[1] https://paperswithcode.com/

Abstractive models are typically more challenging than extractive ones since they create summaries using a vocabulary different from the one used in the input prompt. This task typically requires more training since it involves summarisation and text generation capacities. On the other hand, Extractive methods create summaries by determining the most important words in a text and producing a summary based on them.

Text Summarization's research spans several fields. Initially, this task was deeply associated with Information Retrieval rather than NLP. However, the recent advances in NLP allowed Text Summarization models to improve their results. This diversity of research fields associated with Text Summarization created several taxonomies to classify the work developed in the area. The distinction between extractive and abstractive is the most relevant, but there are others, like single and multi-document summarization models. A multi-document text summarization system uses several documents, usually about the same topic, as input to produce a single summary as output. A single-document model performs a unitary summarization of a single input document [32].

The Text Analysis Conference (TAC) is widely known in this field. Several datasets and architectures are annually proposed in Text Summarization motivated by the many tracks of this conference focused on this NLP Task.

Both summarization methodologies and other information about this task are presented in Section 3.6 of this document.

2.2.3 Language Identification

Language Identification (LID) is the NLP task that detects the language in which an input text is written [20]. Some models are capable of differentiating among varieties of the same languages. This capability of differentiating among variants within the same language is relevant in the Portuguese case to differentiate between European and Brazilian Portuguese. The achievements in this area are documented in Section 3.1 of this document.

2.2.4 Named Entity Recognition

Named Entity Recognition (NER) is the NLP task that identifies entities in a text. NER is a low-level NLP task (2.1.3), serving as a core technique to support higher-level tasks like Question Answering (QA) [68]. NER is also a seq-2-seq task (2.1.4) that maps each token in a corpus to a NER-type label. Most NER datasets follow the BIO labelling scheme used in the CoNLL-2003 shared task [120]. In this labelling scheme, the beginning of an entity is marked with a prefix B, the intermediate tokens with an I, and every token that does not represent an entity is marked with an O (outside) tag.

```
1    U.N.        NNP  I-NP  I-ORG
2    official    NN   I-NP  O
3    Ekeus       NNP  I-NP  I-PER
4    heads       VBZ  I-VP  O
5    for         IN   I-PP  O
6    Baghdad     NNP  I-NP  I-LOC
7    .           .    O     O
```

Listing 2.1: Example of NER annotation in the CoNLL 2003 labelling scheme.

The type of entities identified depends on the specificity of the problem. The CoNLL-2003, the most adopted labelling scheme among NER datasets, includes five types of entities:

- Outside (**O**)
- Persons (**PER**)
- Locations (**LOC**)
- Organizations (**ORG**)
- Miscellaneous Entities (**MISC**)

2.2.5 Semantic Role Labelling

Semantic Role Labelling (SRL) is a low-level NLP task that intends to identify the role of each token in a phrase. Typically defined as the task that identifies "who did what to whom, when and where" [41], SRL has its functionality based on identifying the primary event of a sentence, the predicate. The capability to assign relationships between the predicate and the other tokens of a sentence is a fundamental task for high-level NLP like QA or Machine Translation. Due to its expressiveness, answering "who did what to whom, when and where" [41] defines a intermediate representation of a sentence that high-level NLP models take advantage of to boost their performance [53]. The roles included in an SRL model heavily depend on the application where they are applied [53]. However, most of them are trained in datasets like the Propbank [54, 82] or the Framenet [6], which define generic roles sufficient for many SRL applications.

Section 3.10 of this document covers the research developed for SRL in the Portuguese Language.

2.2.6 Temporal Information Extraction

Temporal Information Extraction task "is the identification of chunks/tokens corresponding to temporal intervals, and the extraction and determination of the temporal relations between those".[2] Temporal Information Extraction is a low-level NLP task used by high-level NLP tasks like QA

[2] https://paperswithcode.com/task/temporal-information-extraction

to integrate timing understanding capabilities. Temporal Information Extraction datasets typically follow an ISO-TimeML standard [90].

```
1 The Navy has changed its account of the attack on the USS Cole in Yemen.
2 Officials <TIMEX3 tid="t1" type="DATE" value="PRESENT_REF" temporalFunction
  ="true" anchorTimeID="t0">now</TIMEX3> say the ship was hit <TIMEX3 tid="t
  2" type="DURATION" value="PT2H">nearly two hours</TIMEX3> after it had
  docked.
3 Initially, the Navy said the explosion occurred while several boats were
  helping the ship to tie up. The change raises new questions about how the
  attackers were able to get past the Navy security.
```

Listing 2.2: Example of temporal extraction annotation in ISO-TimeML format.

2.2.7 Part-of-Speech Tagging

Part of Speech Tagging (POS Tagging) is one of the first low-level NLP tasks where researchers obtained good results. This sequence-to-sequence NLP task maps a group of input tokens to its corresponding part of speech. The progress of POS tagging varies across languages. Seen as a resolved task in Western style languages like English and Portuguese, it remains a challenge for Eastern languages like Mandarin and Japanese, whose language nature is highly supported in the character level which difficult the task of POS tagging model.

POS tagging is one of the most crucial low-level NLP tasks since it serves as the first step to enable the actuation of more high-level models. One example is the incorporation of a POS tagger before performing NER inference in the spaCy ecossytem[3] to boost the overall performance of the NER step.

Traditionally, POS tagging datasets are defined using treebanks. In the case of the English case, the most widely used treebank, The Penn treebank, defines 36 POS tags [117]. More details about this task and its applicability to the Portuguese-specific task are described in this document's Section 3.4.

2.2.8 Relation Extraction

Relation Extraction (RE) is the NLP task that focuses on "finding and classifying semantic relations among entities mentioned in a text, like child-of (X is the child-of Y), or part-whole or geospatial relations" [52]. Relation extraction has close links to knowledge graphs. These graph-based representations help provide a visual picture of the outputs of RE models [52]. Figure 2.4 provides an example of a knowledge graph. The details about how this NLP task is in Portuguese are scrutinised in Section 3.9.

[3] https://spacy.io/

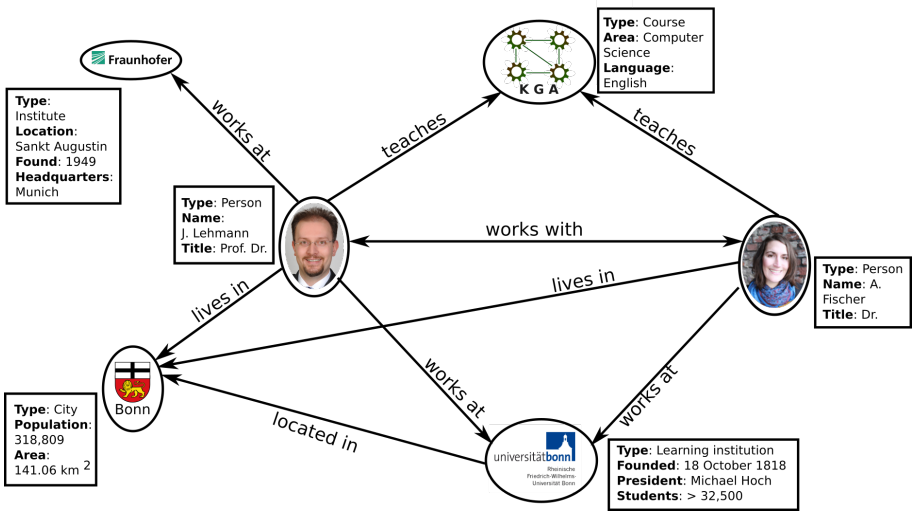


Figure 2.4: Example of knowledge graph built based on relation extraction [109].

2.2.9 Machine Translation

Machine Translation (MT) aims to remove the need for an interpreter to allow two humans not speaking the same language to communicate. This objective is already possible for major languages like English and Spanish. In the Portuguese case, the model's performance remains unclear. It is impossible to easily determine the current benchmarks for this NLP task because the best-performing models are closed-source commercial tools that do not publicly provide any performance metrics of their products.

Not only is MT used by end-users, but it is also a relevant data augmentation tool. An analysis of the commercial and open-source tools, benchmarks and datasets available to train this kind of model is presented in Section 3.2 of this document.

Chapter 3

Portuguese Natural Language Processing

In the previous chapter, we introduced theoretical concepts relevant to ensure the audience understands the concepts covered in this document. This chapter presents the results obtained during our literature review process. We list training processes, datasets, models, the performance achieved and research teams that highlight themselves in eight NLP tasks (2.2). The NLP tasks that will be the object of research in the implementation chapter of this document, NER (5) and Text Summarization (6), witness a more extensive analysis by including additional information about the SOTA architectures for these tasks and an analysis of the relation that commercial libraries, like spaCy, established with them.

3.1 Language Identification

This section focuses on the Language Identification task (LID). We already briefly introduced it in Section 2.2.3 of this document, and now we provide an extensive enumeration of the work available for this field in Portuguese. We specialise in determining if models can differentiate between European and Brazilian Portuguese variants.

3.1.1 Training Process

Language Identification is a text classification supervised task. A statistical LID model learns from a data collection in a specific language [48, 53]. Researchers tend to consider LID a "solved problem" [48], yet, this is only the case for high-resource languages. Identifying variants within the same language or the challenges introduced by multi-language documents still needs to be solved.

3.1.2 Datasets

Many monolingual corpus can be used to train a LID model. Due to this flexibility, there are few datasets specialised in LID. Authors tend to define their corpus themselves before proposing a new LID model. In Table 3.1, we list the corpora introduced in the context of several projects of LID. Then, we describe the results obtained.

Table 3.1: Datasets used in LID projects focused on European (PT) and Brazilian (BR) Portuguese.

Project Name	**Author**	**Year**	**PT Dataset**	**BR Dataset**
Language Identification in Web Pages [65]	Bruno Martins, Mário J. Silva	2005	Web Crawled Data	-
Identification of Document Language is not yet a completely solved problem [27]	Joaquim Ferreira da Silva	2006	Eurolex News from Publico Diário de Notícias Newspaper	Legal Documents from the General Secretariat of Brazilian Republic Presidency
Automatic Identification of Language Varieties: The Case of Portuguese [128]	Marcos Zampieri	2012	News from Diário de Notícias	News from Folha de São Paulo
Smoothed n-gram based models for tweet language identification: A case study of the Brazilian and European Portuguese national varieties [17]	Dayvid W. Castro	2017	Portuguese Tweets published with a Portuguese IP address	Portuguese Tweets published with a Brazilian IP address

Language Identification in Web Pages [65] : Used a web crawler to aggregate corpora in twelve languages. This project differentiates from the remaining ones since it does not focus on identifying Portuguese variants.

Identification of Document Language is Not yet a Completely Solved [27] : Introduces European and Brazilian corpora to distinguish between those two Portuguese variants. The European corpus was composed of a combination of news from the Portuguese newspapers Público and Diário de Notícias and juridical corpora from the European Court of Justice. The Brazilian variant was composed of legal documents from the Brazilian ministry of education and a web newsletter from the Brazilian Republic office.

Automatic Identification of Language Varieties: The Case of Portuguese [128] : Collection of journalistic corpus to address the challenge of distinguishing between European and Brazilian Portuguese variants. The European Portuguese corpus is composed of news from Diário de Notícias, and the Brazilian is composed of news from the Folha de São Paulo newspaper.

Smoothed n-gram based models for tweet language identification [17] : Addressed the distinction between European and Brazilian Portuguese variants by training a model on a compilation of tweets. The research team used the tweet's location as a heuristic to obtain an unsupervised mechanism to collect Portuguese variant sensitive data automatically.

3.1.3 Models

We notice differences between the SOTA models used for multi-language identification and the research we highlight in this document. The multi-language SOTA models are typically introduced by big IT companies that require powerful LID models to differentiate among hundreds of different languages spoken around the globe. These models are based on neural-based methods.

As described in Table 3.3, none of the four more popular tools for LID published publically by multinational big tech companies distinguish between European and Brazilian Portuguese. Our research focused, however, on the Portuguese case, and its European and Brazilian variants showed that this particular task could be handled using less complex language modelling schemes. The results for the Portuguese case are listed in Table 3.2.

Table 3.2: Summary of the performance of LID models that focus on European (PT) and Brazilian (BR) Portuguese cases. Some test sets are defined by the authors(A.D).

Project Name	Author	Year	Test Set	Methodology	Results PT	Results BR
Language Identification in Web Pages [65]	Bruno Martins	2005	A.D	8 n-grams	0.920 Precision	
Identification of Document Language is not yet a completely solved problem [27]	Joaquim Ferreira Da Silva	2006	A.D	Quadratic Discrimination Score and character n-grams (from 2 to 8)	0.986 Precision	0.987 Precision
Automatic Identification of Language Varieties: The Case of Portuguese [128]	Marcos Zampieri	2012	A.D	Log-Likelihood and character 4-grams	0.998 Accuracy	
Smoothed n-gram based models for tweet language identification: A case study of the Brazilian and European Portuguese national varieties [17]	Dayvid W. Castro	2017	A.D	Naïve Bayes ensemble model combined with TF-IDF	0.934 Precision	0.912 Precision

Language Identification in Web Pages [65] : Created a model based on 8 grams to differentiate between European and Brazilian Portuguese web pages.

Identification of Document Language is not yet a completely solved problem [27] : Tested the impact of models based on different n-grams to test LID capabilities among European and Brazilian Portuguese corpora.

Automatic Identification of Language Varieties: The Case of Portuguese [128] : N-Gram based strategy to differentiate among European and Brazilian Portuguese variants.

Smoothed n-gram based models for tweet language identification [17] : Trained a Naive Bayes classifier to determine the variant in Portuguese tweets.

Table 3.3: Analysis of commercial LID models. Focusing on languages (L.C) and the Portuguese variants covered.

Project Name	Author	Year	Methodology	L.C	Portuguese Variants
Fast Text	Meta	2017	Embeddings	176	No Distinction
Language Identification from Very Short Strings	Apple	2019	Bi-LSTMs	n/a	n/a
Compact Language Detector v3 (CLD3)	Google	2020	Embeddings	107	No Distinction

3.1.4 Results

Table 3.2 summarises the results observed for the models focusing on the Portuguese language case. The scores obtained by different research teams are above 90% precision using N-Gram-based methodologies. These figures show that it is possible to identify Portuguese variants with extraordinary performance.

3.2 Machine Translation

This section describes Machine Translation (MT) systems. This NLP task was briefly introduced in this document's Section 2.2.9. Here, we describe the SOTA resources for MT in the Portuguese case. Due to MT's particular commercial relevance, several SOTA solutions are offered by big IT companies like Microsoft, Google or Amazon, oriented for the end users. In order to easily distinguish open academic research from closed-source MT research, we decided to present these results in separate subsections.

3.2.1 Training Process

Machine Translation is a supervised sequence-to-sequence NLP task. MT models' goal is to map tokens in the original language in the target one [53]. As mentioned in Section 2.2.1 of this document, MT systems usage is not restricted to commercial usage by end-users but is also

a powerful DA technique. However, applying MT in the context of DA introduces additional challenges; the need to perform label alignment is one of them [110, 47]. Due to the heterogeneity of human languages, the first token of an English sentence could become the last token in Japanese. The training of MT systems is, therefore, the capability of learning to establish correct linking of the different tokens of an input sentence.

English: *He wrote a letter to a friend*
Japanese: *tomodachi ni tegami-o kaita*
friend to letter wrote

Figure 3.1: Example of the different token alignment in English and Japanese language [53].

3.2.2 Machine Translation Evaluation Metric: BLEU

Evaluating the performance of MT models is challenging due to the wide range of good-quality translations possible for a simple input prompt. To mitigate this problem, BLUE was introduced in 2002 [83].

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} w_n \cdot \log p_n\right) \tag{3.1}$$

Where:

$$\text{BP} = \begin{cases} 1 & \text{if } c > r \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r \end{cases} \tag{3.2}$$

BLEU stands for Bi-Lingual Evaluation Understudy and is a modified precision evaluation metric to accommodate the particularities of MT. It uses a sum of partial token occurrence counts, typically from uni-grams to four-grams, as the factor of a monotonic exponential function [71]. The BLEU score includes a Brevity penalty (BP) to penalise shorter sentences that tended to obtain a higher BLEU due to the mathematical formulation of this metric.

3.2.3 Datasets

Training MT models require large volumes of parallel translations of the same sentence in the set of languages covered by the MT model [39]. Modern MT models require "tens or even hundreds of millions of parallel sentences" [39] in their training process. This fact implies that MT datasets are typically extensive collections of aligned data.

Jörg Tiedemann, in association with the Nordic Language Processing Laboratory and the University of Helsinki, launched in 2003 the Open Parallel Corpus(OPUS) project [119]. OPUS is a widely used platform in the context of MT that catalogues several parallel corpora in 727 languages. The datasets introduced by this project are in XCES format, an XML-based approach that uses a floating point to establish the alignment between parallel corpus in different languages. An

example of a parallel XCES corpus between European and Brazilian Portuguese is provided in Figure 3.2.

```
- <s id="s1">
    <w id="w1.1">O</w>
    <w id="w1.2">pacote</w>
    <w id="w1.3">%</w>
    <w id="w1.4">s</w>
    <w id="w1.5">versão</w>
    <w id="w1.6">%</w>
    <w id="w1.7">s</w>
    <w id="w1.8">tem</w>
    <w id="w1.9">uma</w>
    <w id="w1.10">dependência</w>
    <w id="w1.11">não</w>
    <w id="w1.12">satisfeita</w>
    <w id="w1.13">:</w>
  </s>
```

```
- <s id="s1">
    <w id="w1.1">O</w>
    <w id="w1.2">pacote</w>
    <w id="w1.3">%</w>
    <w id="w1.4">s</w>
    <w id="w1.5">versão</w>
    <w id="w1.6">%</w>
    <w id="w1.7">s</w>
    <w id="w1.8">tem</w>
    <w id="w1.9">uma</w>
    <w id="w1.10">dependência</w>
    <w id="w1.11">desencontrada</w>
    <w id="w1.12">:</w>
  </s>
```

Figure 3.2: Example of corpus catalogued by OPUS in the XCES format [119].

Due to the relevance of OPUS in this NLP field, the XCES format has established itself as a standard. It is common to find recent MT corpus in this format. Our research identified seven datasets. The information collected is summarized in Table 3.4 and described below.

Table 3.4: Analysis of MT European (PT) and Brazilian (BR) Portuguese datasets.

Project Name	**Author**	**Year**	**PT Dataset**	**BR Dataset**
EUROPARL Corpus [55]	Philipp Koehn	2005	European Parliament Transcripts	-
Open Subtitles [61]	Jörg Tiedemann	2016	Movies Subtitles in European Portuguese	Movies Subtitles in Brazilian Portuguese
Tilde MODEL Corpus Multilingual Open Data for European Languages [97]	Roberts Rozis	2017	Multiple European Union website data published in European Languages	-
QTLeap News Corpus	António Branco	2016	n/a	n/a
Lidioms [70]	Diego Moussallem	2018	-	Web Crawled Data
Wiki Matrix [104]	Facebook	2019	Wikipedia European Parliament	Wikipedia
Law Health Corpus for Translation	Gabriel Lopes	-	n/a	n/a

EUROPARL Corpus Parallel Corpora: Portuguese-English [55] : This dataset aligned the content transcriptions made during the public sessions at the European Parliament. This information

is publicly available by the European Commission in the several languages spoken in the European Union(EU) but requires indexation and alignment. It is a corpus composed exclusively of European Portuguese documents. This dataset is available under the OPUS platform.

QTLeap News Corpus : A parallel corpus for Basque, European Portuguese, Dutch; Flemish, Bulgarian, English, German, Czech, Spanish; Castilian developed by the QTLeap research team(Quality Translation by Deep Language Engineering Approaches) at the University of Lisbon. This dataset is available on the platform PortugalClarin[1].

OpenSubtitles [61] : Corpus introduced by the OPUS project and covers subtitles of movies in 62 different languages conveniently aligned in the XCES format. This dataset distinguishes between European and Brazilian Portuguese.

Tilde MODEL Corpus – Multilingual Open Data for European Languages [97] : Tilde is a parallel corpus focused on European languages. It catalogues European Portuguese content from several European web resources.

LIdioms - A Multilingual Linked Idioms Data Set [70] : Dataset introduced in the context of the Language Resources and Evaluation Conference(LREC) 2018. This dataset offers a multilingual RDF representation of sentences in "five languages: English, German, Italian, Portuguese, and Russian" [70].

WikiMatrix [104] : Parallel corpus provided by Facebook. This dataset provides Wikipedia-aligned text for 85 different languages. It does not distinguish between different variants of Portuguese, including European and Brazilian Portuguese corpora.

Law Health Corpus for Translation : Parallel corpus including European Portuguese in the Law and Health domains by Gabriel Lopes, principal research at Universidade Nova de Lisboa. This dataset is available on the platform PortugalClarin.

To conclude, there are some parallel corpora available for the Portuguese case. OPUS [119] did an excellent job cataloguing datasets for this particular NLP task. Therefore, the best takeaway from this document's section is to consult OPUS as the primary source of linguistic resources for MT when faced with this type of NLP problem.

3.2.4 Open Source Models

We identified two projects focused on open-source MT for the Portuguese case. The results are summarised in Table 3.5 and described below.

LX-Translator [101] : European Portuguese-Chinese MT system developed at NLX-Natural Language and Speech Group of the Department of Informatics at the University of Lisbon. This research trained a Transformer based model using the Marian framework. Marian is "an efficient Neural Machine Translation framework written in pure C++ with minimal dependencies" [49]

[1] https://portulanclarin.net/

Table 3.5: Analysis of Open Source MT models that focus on Portuguese language (Lang).

Project Name	Author	Year	Test Set	Lang. Pairs	BLEU
LX-Translator [101]	António Branco	2019	UMPCorpus	Portuguese-Chinese	PT-ZH 13.98 ZH-PT 16.23
Lite Training Strategies for Portuguese-English and English-Portuguese Translation [62]	Alexandre Lopes	2020	Adapted ParaCrawl	Portuguese-English	PT-EN 45.99 EN-PT 38.12

Lite Training Strategies for Portuguese-English and English-Portuguese Translation [62] : In this project, researchers explored the T5 transformer model [92] to perform MT under low hardware resources conditions.

Despite being one of the most used NLP Tasks, there is little recent academic research in MT focused on the Portuguese case. From the literature we consulted, the resources required to train a SOTA MT are incompatible with the resources available in the Portuguese language. In addition, MT is an NLP task monopolised by bit IT companies that offer easy access to MT models that outscore any academic approach to this problem in Portuguese due to the high amount of linguistic and hardware resources these companies have.

3.2.5 Commercial Models

As introduced previously in this document, MT is an NLP task dominated by the usage of commercial MT models provided by big IT companies like Google with its Google Translator and Google Cloud Services, Microsoft (MS) with its Bing Translator and Azure Translation Services and Amazon Web Services (AWS) and its Amazon Translation services. Whereas Google Translate and Bing Translate are free services oriented to the end user, MS Azure and AWS Translation are paid services that offer more sophisticated IT-oriented services. These tools provide solutions based on APIs that simplify the integration of MT capabilities within other software components.

Commercial services do not offer transparent metrics about their current performances, the data where their models are trained and which architectures they use. These services are, therefore, complete black boxes to the community. In Table 3.6, we describe the functionalities offered by these services. With a free subscription for students of 2 million translated tokens, Amazon Translate offers a batch translation service, while Azure and Google expose a REST API to perform translation requests to the service. Both MS Azure and AWS differentiate between European and Brazilian Portuguese.

3.2.6 Results

Table 3.5 lists the results of our research for MT in the context of the Portuguese Language. The LX-Translator project [101] scores poor BLEU scores while translating corpora to Chinese. The

Table 3.6: Analysis of Commercial MT Models. Focusing on the European(PT) and Brazilian(BR) variants of Portuguese language

Tool Name	PT	BR	Batch Translation	API	Free Subscription
Amazon Translate	Y	Y	Y	N	2M/Month
Azure Translate	Y	Y	N	Y	2M/Month
Google Cloud Services Translation	N	Y	Y	Y	N

Lite Training Strategies for Portuguese-English and English-Portuguese Translation project [62] achieved a higher BLEU score than the current SOTA BLEU scores obtained by many open-source transformer-based models in English language [2].

Unfortunately, commercial MT engines provided by big IT companies do not provide transparent information regarding their models' BLEU scores. We aim to understand better the current SOTA benchmarks for these tools in Section 6.2.

3.2.7 Research Teams

Our survey identified the OPUS project headed by Jörg Tiedemann at the Language Technology lab at the University of Helsinki as a significant contribution to catalogue data sources of European languages to support MT development.

3.3 Named Entity Recognition

This section describes the current SOTA for Named Entity Recognition(NER) systems. This NLP task was briefly introduced in this document's Section 2.2.4. Here, we focus our attention on describing the Portuguese NER research.

3.3.1 Training Process

Named Entity Recognition is a sequence-to-sequence NLP task. NER models map input natural language tokens to their corresponding NER labels. Therefore, this NLP task is a supervised learning task since it requires a previously annotated dataset of this mapping between text and entities.

3.3.2 The F1-Score Evaluation Metric

Several performance metrics are available to quantify the quality of a NER engine; however, F1-Score is usually used in literature to serve this purpose. F1-Score is "the harmonic mean of precision and recall" [56]. In this dissertation, we present NER results using F1-Score to simplify the results' comparability with other projects.

[2] https://paperswithcode.com/sota/machine-translation-on-wmt2014-english-german

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{3.3}$$

Its formula (3.3) implies that a higher value of F1-Score forces the NER model to present a higher combination of precision and recall; this forces the model not only to correctly classify the entities identified but also to identify the correct number of elements in the text.

3.3.3 Datasets

We were able to identify eight Portuguese NER datasets. Due to the relevance of this NLP task, NER is one of the tasks addressed in this document that we were able to catalogue more resources. The conclusions of this section are summarised in Table 3.7 followed by a description.

Table 3.7: Analysis of Portuguese NER dataset properties. Including information regarding the Portuguese variants of the corpus (Variant)

Project Name	Author	Year	Corpus Style	BIO Format	Variant
Primeiro HAREM [98]	Diana Santos	2005	Journalistic Text	N	BOTH
Mini-HAREM [102]	Luís Sarmento	2007	Journalistic Text	N	BOTH
Segundo HAREM [37]	Cláudia Freitas	2008	Journalistic Text	N	BOTH
CINTIL [7]	António Branco	2012	Journalistic Text Legal Documents	Y	PT-PT
WikiNER [74]	Joel Nothman	2012	Wikipedia	Y	BOTH
SIGARRA News Corpus [86]	André Pires	2017	Newsletters	N	PT-PT
LeNER-Br [64]	Pedro Araujo	2018	Legal Documents	Y	PT-BR
1758 Portuguese Parish Memories [126]	Renata Viera	2021	Historical Letters	Y	PT-PT
Wikineural [118]	Simone Tedeschi	2021	Wikipedia	Y	BOTH
MAPA [4]	E. Ajausks	2022	Legal Documents	Y	PT-PT

HAREM (First, Second and Mini HAREM) [69] : This dataset combines European and Brazilian annotated data in XML format. It is the most used dataset for benchmarking Portuguese NER. It compresses three variations, the First [98], Second [37] and Mini HAREM [102]. None of these HAREM datasets follows the conventions established by the ConLL-2003 dataset regarding file structures and the labels used during this conference.

CINTIL Corpus-Corpus Internacional do Português [7] : European Portuguese corpus that provides not only NER annotations in the CoNLL-2003 format but also POS tagging and Lemmatization labels.

WikiNER [74] : Multilingual NER dataset based on Wikipedia information that catalogues data from both European and Brazilian Portuguese. It is provided in the BIO format.

SIGARRA News Corpus [86] : Manually annotated NER corpus based on SIGARRA, the University of Porto's internal system, that provides several administrative features, including a weekly newsletter, the object of the labelling process.

LeNER-Br [64] : NER dataset of legal Brazilian Portuguese documents. It follows the standard convention established by ConLL conferences shared tasks.

Enriching the 1758 Portuguese Parish Memories (Alentejo) with Named Entities [126] : A manually annotated dataset from historical European Portuguese documents.

Wikineural : NER dataset based on data from Wikipedia covering nine languages, including Portuguese. Our analysis verified that this dataset includes European and Brazilian data and is provided in ConLL standard format.

MAPA [4] : A European Project focuses on anonymization, providing a multilingual legal documents dataset for different NLP tasks, including NER. It is provided in the BIO format.

3.3.4 Models

Our literature review identified eight projects that offer Portuguese NER models. The results are presented in a compact form in Table 3.8 followed by a description.

CRPC-Named Entity Recognizer : NER model trained in the European Portuguese CINTIL corpus [7] and developed by the Center of Linguistics of the University of Lisbon.

Named entity extraction from Portuguese web text [86] : This project used a combination of HAREM and the SIGARRA as data sources to train Portuguese NER models. The author published its contributions to spaCy[4] and OpenNLP[5], two popular NLP frameworks, simplifying the inter-compatibility of the models produced.

Portuguese Named Entity Recognition using BERT-CRF [115] : Research focused on adapting the Portuguese BERT, BERTimbau [114, 113] to NER, using a CRF decoder to fine-tune the architecture on First HAREM [69].

Contributions to Clinical Named Entity Recognition in Portuguese [63] : This project focused its attention on clinical corpora and applied a "Bidirectional Long Short Term Memory with a stacked Conditional Random Fields layer (BiLSTM-CRF)" [63] to a private corpus provided by the Coimbra University Hospital Centre (CHUC).

[4] https://spacy.io/
[5] https://opennlp.apache.org/

Table 3.8: Analysis of Portuguese NER models With F1-Score performance for European (PT-F1) and Brazilian (BR-F1) Portuguese. Some test sets are defined by the authors (A.D)

Project Name	**Author**	**Year**	**Methodology**	**Test Set**	**PT-F1**	**BR-F1**
CRPC Named Entity Recognizer[3]	Amália Mendes	-	n/a	CINTIL	0.979	-
Named entity extraction from Portuguese web text[86]	André Pires	2017	spaCy, OpenNLP	SIGARRA Corpus	0.866	-
Contributions to Clinical Named Entity Recognition in Portuguese [63]	Fabio Lopes	2019	BiLSTM-CRF	A.D	0.739	-
IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks [22]	Sandra Collovini	2019	BiLSTM-CRF-FlairBBP	A.D	0.650	
Assessing the Impact of Contextual Embeddings for Portuguese Named Entity Recognition [99]	Joaquim Santos	2019	FlairBBP + Word2Vec Embeddings	HAREM	0.823	
***Portuguese Named Entity Recognition using BERT-CRF*[115]**	**Fabio Souza**	2019	**BERT + CRF**	HAREM	**0.832**	
Enriching the 1758 Portuguese Parish Memories (Alentejo) with Named Entities [126]	Renata Vieira	2021	spaCy, BCF	A.D	0.458	-
A Biomedical Entity Extraction Pipeline for Oncology Health Records in Portuguese [112]	Hugo Sousa	2023	BERT+UMLS Entity Linking	A.D	-	0.842
spaCy Portuguese NER	Explosion	-	Transition-Based Parser Model	WikiNER	0.900	-

IberLEF 2019 Portuguese NER and Relation Extraction Tasks [22] : Overview of the results obtained by participants in the IberLEF 2019-Iberian Languages Evaluation Forum' on Portuguese NER and RE tasks.

Assessing the Impact of Contextual Embeddings for Portuguese NER [99] : The author studied the performance of Brazilian NER models based on word embeddings on two different datasets. They achieve SOTA benchmarks on the NILC Embeddings dataset.

Enriching the 1758 Portuguese Parish Memories (Alentejo) with Named Entities [126] : This

project focus on evaluating the performance of existing NER models on a historical corpus manually annotated by the authors. They used spaCy and BiLSTM-CRF [100], a model based on the combination of LSTMs and Conditional Random Fields(CRF) to identify entities in corpora.

A Biomedical Entity Extraction Pipeline for Oncology Health Records in Portuguese [112] : Finetuned BERTimbau with post-processing stage of entity linking in a dataset manually annotated by doctors at IPO Porto.

spaCy Portuguese NER Models : It is not clear what is the current underlining architecture used by the spaCy NER model. There are short references to a "transition-based parser model", and older spaCy versions used CNN-based neural networks. Additionally, spaCy major version 3 introduces the possibility of using transformer-based NER models. However, they are not yet available in the Portuguese language. Explosion, the company that manages spaCy, clarifies that current Portuguese NER models are trained in the WikiNER dataset [74]. This corpus includes European and Brazilian sources. Therefore spaCy does not distinguish between European and Brazilian variants of Portuguese.

3.3.5 Results

The results obtained for the Portuguese case are listed in Table 3.8. The results demonstrate a high variance in the F1-Scores obtained heavily dependent on the test set used. Additionally, Conditional Random Fields(CRF) based models are the technique that provided the best results in the research identified.

The variance in the results obtained is evident in projects that use spaCy. While the official spaCy document publishes a performance of 0,90 F1-Score, Renata Vieira et al. [126] obtained 0,06 F1-Score for the Organization tag; while André Pires [86] obtained 0,40 F1-Score using another test set. To address this uncertainty regarding the actual spaCy performance, we present in Section 5.6 a pipeline to extract NER benchmarks using this tool.

3.3.6 Research Teams

During our literature review of NER resources for the Portuguese language, we identified two significant contributions to assist the research on this NLP task. The first is David Baptista's GitHub[6] that extensively catalogues NER datasets in several European languages, including Portuguese. The second is a contribution of the project previously identified headed by Joaquim Santos at PUCRS [99]. This research team published their work in a GitHub repository[7] that provides links to several resources for Brazilian NER.

[6] https://github.com/davidsbatista
[7] https://github.com/jneto04/ner-pt

3.4 Part Of Speech Tagging

This section describes the current SOTA for POS Tagging in the Portuguese Language. POS tagging was already briefly described previously in Section 2.2.7 of this document. This section describes the SOTA training process for POS, the datasets and models used, and the benchmarks obtained by those models.

3.4.1 Training Process

Part of Speech tagging is a sequence-to-sequence supervised NLP Task. POS engines map the input tokens to their corresponding POS label. This NLP task is similar to the NER one described in Section 3.3. The consequence is that most NER datasets also include POS tagging resources and can be used for both purposes. As mentioned before (2.2.7), POS tagging typically precedes NER. The spaCy library uses POS Tagging information to improve the results obtained by NER models.

3.4.2 Datasets

Our literature review for POS tagging was done posteriorly to revising the NER task. As already mentioned, NER datasets typically include POS tagging features. Since we could identify eight NER Portuguese datasets from diverse sources, we split the time dedicated to POS Tagging resources in two. Firstly, we determine if each NER dataset includes POS Tagging features (3.9). Secondly, we survey the existence of Portuguese datasets focused exclusively on POS Tagging (3.10).

Table 3.9: Analysis of POS Tagging properties of Portuguese NER datasets.

Project Name	**POS Tagging**
HAREM	N
Wikineural	N
CINTIL	Y
SIGARRA News Corpus	N
LeNER-Br	N
1758 Portuguese Parish Memories	N
WikiNER	Y

Our literature review identified the project Floresta Sintática [3] as the primary source of POS tagging labelled data for Portuguese. This project includes POS tags for two Portuguese journalistic corpora, the CETEMPúblico for European Portuguese and the CETENFolha for Brazilian Portuguese. Table 3.10 synthesises the information collected.

Table 3.10: Analysis of POS Portuguese datasets including information regarding Portuguese variants.

Project Name	Authors	Year	BIO Format	Text Type	Variants
CETEMPúblico	Linguateca	2008	N	Journalistic Text	PT-PT
CETENFolha	Linguateca	2008	N	Journalistic Text	PT-BR

3.4.3 Models

Our research identified two Portuguese POS taggers. The number of resources available is scarce. Our hypothesis for this scarsity is that POS Tagging for Western languages is already a "solved" problem. Therefore there is no point in creating more tools to address a solved problem. The information collected is synthesised in Table 3.11 and described below.

Table 3.11: Analysis of Portuguese POS Taggers. Focusing on the training dataset (T.D) and its Portuguese variant (Variant)

Project Name	Authors	Year	T.D	Methodology	Test Set	Accuracy
LX-Tagger [10]	António Branco	2006	CINTIL-Propbank	Brills TBL	CINTIL-Propbank	0.970
spaCy	Explosion	n/a	Floresta Sintática	n/a	Floresta Sintática	0.970

LX-Tagger : Tool developed by António Branco at the University of Lisbon using the Brills TBL algorithm [10].

spaCy Portuguese POS Tagger : This model was trained in the Floresta Sintática [3] corpus that includes European and Brazilian Portuguese corpora. spaCy does not clarify the architecture used in its POS Tagger.

3.4.4 Results

The results presented in Table 3.11 confirm that POS tagging is a "solved" NLP task. The results presented have incredibly high levels of accuracy. However, due to the lack of transparency of spaCy and the outdated interface of LX Tagger, some work in this field can be done to improve the results' completeness.

3.5 Temporal Information Extraction

This section presents the resources identified for Portuguese Temporal Information Extraction(TIE). This NLP task introduces models capable of retrieving temporal information from corpora. This

task was briefly described in Section 2.2.6. Temporal Information Extraction increased in relevance after the SemEval 2013 conference hosted a TIE-shared task. The work at SemEval introduced the first TIE benchmark dataset, the TempEval-3 [122]. Here we list the resources available for this NLP task in Portuguese.

3.5.1 Training Process

Temporal Information Extraction is a supervised sequence-to-sequence task. A statistical TIE model maps the input tokens towards labels that bind the beginning and end of temporal expressions. Rule-based solutions are also often applied to solve this NLP task.

3.5.2 Datasets

We identified a single Portuguese TIE dataset, the TimeBankPT [25]. This dataset is based on TimeML [90], an English TIE dataset used as the primary benchmarking source. TimeBankPT was automatically translated using Google Translator with posterior manual adjustments, labelling alignment and corrections by human annotators. Additionally, this dataset implements the ISO standard of TIE annotation introduced by TimeML (2.2.6).

We present two tables that synthesise the research done. First, we introduce a Table 3.12 that summarises the datasets found for the Portuguese Case. Then, due to the low number of resources in Portuguese, we decided to introduce an additional second table (3.13) that lists the TIE datasets we identified in the English Language.

Table 3.12: Analysis of European(PT) and Brazilian(BR) Portuguese TIE datasets.

Project Name	**Authors**	**Year**	**PT Dataset**	**BR Dataset**
TimeBankPT [25]	António Branco	2012	TimeML Translation	-

Table 3.13: Analysis of English TIE datasets.

Project Name	**Authors**	**Year**	**Corpus Type**
TimeBank1.2 [89]	James Pustejovsky	2006	Journalistic Text
TempEval-3 [122]	Naushad UzZaman	2013	Multiple Sources

3.5.3 Models

Our survey identified three TIE Portuguese models; the results are summarized in Table 3.14 and described below.

LX-TimeAnalyzer [24] : Decision Tree based TIE tagger.

Table 3.14: Analysis of F1-Score performance in European(F1-PT) and Brazilian(F1-BR) of Portuguese TIE models.

Project Name	Author	Year	Test Set	F1-PT	F1-BR
Portuguese HeidelTime [116]	Jorge Mendes	-	Portuguese TimeBank 1.0	0.720	-
LX-TimeAnalyzer [24]	António Branco	2012	TempEval-2	0.800	-
Tieval [111]	Hugo Sousa	2023	TempEval-3	0.812	-

HeidelTime TIE tagger [116] : This project aims to create a cross-linguistic framework to establish baselines for TIE in all major languages. Jorge Mendes published the Portuguese version with the supervision of Ricardo Campos at Politécnico de Tomar.

Tieval : Ensemble of preexisting TIE tools in different languages.

3.5.4 Results

The results obtained for the Portuguese case, summarized in Table 3.14, demonstrate good F1-Scores. However, the low number of resources available introduces the need for further validation with different data sources. Introducing new benchmarking corpora would guarantee the correctness of the values already existent.

3.5.5 Research Teams

We identified two stakeholders associated with the Portuguese TIE task during our survey. Ricardo Campos at the Politécnico de Tomar. Ricardo is an affiliated researcher of the European project Heideltime, one of the significant stakeholders in TIE in Europe. Moreover, we highlight the work developed in Lisbon at the FCUL headed by António Branco. Together these two research teams provide the resources available to assess this task in Portuguese NLP.

3.6 Text Summarization

We briefly introduced Text Summarization previously in this document (2.2.2). We mentioned that there are two major types of text summarization, extractive (3.7) and abstractive (3.8). In this section, we introduce the datasets, models, results obtained and the research teams working on text summarization for the Portuguese case.

3.6.1 Extractive and Abstractive Text Summarization

In extractive text summarization (ETS), the words used in the resulting summary are constrained to the ones included in the input document [32]. In abstractive text summarization (ATS), the

summary can include words not part of the original corpus. For this reason, ATS requires more advanced Text Generation capabilities to produce a correct summary.

Text Summarization is a field that spans Information Retrieval(IR) and NLP. At the same time, it is an IR and an NLP task. Abstractive summarization is more connected with NLP than ETS due to its text generation needs [32]. This document covers an NLP perspective of Text Summarization; for this reason, we focus our attention mainly on ATS.

3.6.2 Single and Multi-Document Summarization

In addition to categorizing extractive and abstractive, text summarization also has an essential distinction between Single and Multi-document summarization. Multi-document summarization produces a single summary based on a batch of documents(batch of emails/news/tweets), while a single document produces a summarization of a single unit of text [32]. According to H. Lin and V. Ng [60], there needs to be more multi-document summarization research. The author noticed that most of the research in this field is focused on the single document case.

The following sections cover the resources, datasets, models and results for both ATS e ETS in Portuguese. Additionally, we created two sections to mark a clear division between both techniques.

3.6.3 Text Summarization Model's Generic Architecture

Text Summarization models, either ETS or ATS, follow a generic architecture composed of the following tasks [32].

1. **Pre-Processing:** Generate a structured representation of the original text using a Language Model [32]. N-grams and word embeddings are some of the most used language modelling techniques in a text summarization task [11].

2. **Processing:** Apply one of the Summarization techniques to identify the most relevant parts of the corpus

3. **Post-Processing:** Use a text generative technique to translate the most relevant parts of the corpus identified into a human-readable summary.

The difference between the many Text Summarization techniques can be seen as variations in each of the items listed above. This way, the primary differentiation between ETS and ATS can be observed as a variation in the Post-Processing phase.

3.6.4 The ROUGE Evaluation Metric

Evaluating the performance of a Text Summarization model is a complex task. The concept of good summarization is ambiguous and heavily dependent on the most valued criteria in each summarization project [72]. Before ROUGE-based metrics were introduced, manual evaluation was the primary source of benchmarks in this NLP task [59].

ROUGE is the current SOTA evaluation metric for text summarization. It is a recall-based metric that calculates the n-gram overlapping between the summary obtained by the model and the summaries that compose the test set [57, 59]. The ROUGE metric has three variations, ROUGE-1, ROUGE-2, and ROUGE-L. The first case measures the overlapping between unigrams; the second is between bigrams; ROUGE-L, the metric that we will use to evaluate our ATS models, considers the overlapping between the longest common subsequence (3.4).

$$\text{ROUGE-L} = \frac{\text{Longest Common Subsequence (LCS)}}{\text{Length of the expected summary}} \tag{3.4}$$

$$\text{LCS}(X,Y) = \begin{cases} 0 & \text{if length}(X) = 0 \text{ or length}(Y) = 0 \\ \text{LCS}(X-1,Y-1)+1 & \text{if last element}(X) = \text{last element}(Y) \\ max(\text{LCS}(X,Y-1),\text{LCS}(X-1,Y)) & \text{if last element}(X) \neq \text{last element}(Y) \end{cases} \tag{3.5}$$

We selected ROUGE-L since many projects evaluate their ATS systems in this metric; this makes comparability easier, and, when compared with ROUGE-1 and ROUGE-2, it offers more correctness of what a good summary is in practice.

However, traditional ROUGE is unsuitable for correctly evaluating the summaries obtained by ATS tools [72]. The nature of ATS is to provide a summary with words that are not constrained to the vocabulary of the original corpora, something that is contrary to the concept of n-gram overlapping. This limitation prejudicates the rigorous evaluation of ATS systems.

Although the cons of using ROUGE, the absence of viable solutions with simple implementation turns this metric widely used to evaluate Text Summarization models [72], all the authors consulted are unanimous that new evaluation metrics are required in order to improve the effectiveness of the model training process. Jun-Ping Ng et.al [72] propose an improved version of ROUGE for ATS systems. The authors successfully used word embeddings in combination with Spearman and Kendall rank correlation analysis to better determine the proximity between the obtained summary and the test set.

3.6.5 Datasets

In order to effectively train and rigorously evaluate the quality of the summaries, it is common to use extractive datasets to create extractive systems and abstractive summaries to train an abstractive model. Nevertheless, this is not a hard requirement for a text summarization system. We identified twelve datasets focused on Text Summarization in the Portuguese Language. In our research, we explicitly differentiate the extractive/abstractive and single/multi-document nature of these datasets. The results of the survey are summarized in Table 3.15 and described below.

Table 3.15: Analysis of Portuguese text summarization datasets Properties. Including information regarding the Type of Corpus (C.T) included, if it is an Abstractive (A) or Extractive (E) dataset, and if it is a Single Document (S.D) or Multi-Document (M.D) dataset, along with identifying the Portuguese Variant (Variant) of the corpus.

Project Name	**Author**	**Year**	**C.T**	**A.T/E.T**	**S.D/M.D**	**Variant**
TeMário [84]	Thiago Alexandre Salgueiro Pardo	2003	Newspapers	A	Single	PT-BR
CM3News [121]	Fabricio E. da S. Tosta	2013	Newspapers	A	Multi	PT-BR
Priberam Compressive Summarization Corpus [5]	Miguel B. Almeida	2014	Newspapers Articles	A	Multi	PT-PT
Summ-it [21]	Sandra Collovini	2016	Newspaper	A	Single	PT-BR
CSTNews-Update [14]	Paula C. F. Cardoso	2017	Newspapers Articles	A	Multi	PT-BR
RulingBR [124]	Diego de Vargas Feijó	2018	Brazilian Supreme Court Decisions	A	Single	PT-BR
A New Annotated Portuguese/Spanish Corpus for the Multi-Sentence Compression Task [87]	Elvys Linhares Pontes	2018	Brazilian Google News	A	Single	PT-BR
WikiLingua [33]	Faisal Ladhak	2020	Wiki How	A	Single	PT-BR
A Corpus for Sentence Compression [73]	Fernando Antônio Asevedo Nóbrega	2020	G1 Newspaper	E	Single	PT-BR
XL-Sum [40]	Tahmid Hasan	2021	BBC News	A	Single	PT-BR
BrWac2Wiki [76]	Andre Seidel Oliveira	2021	Wikipedia	A	Multi	PT-BR
OpiSums-PT [23]	Marcio Lima Inácio	2021	Books and Reviews	BOTH	Single	PT-BR

TeMário: Um Corpus para Sumarização Automática de Textos [84] : This Brazilian Portuguese single-document abstractive dataset includes one hundred news articles whose summary was manually annotated.

CM3News [121] : Multi-document abstractive summarization corpus in Brazilian Portuguese includes several documents in ten subjects. The introduction of these ten clusters of documents constitutes the support for multi-document text summarization. The documents were collected from several Brazilian news outlets.

Priberam Compressive Summarization Corpus [5] : This dataset introduced by the company Priberam is a multi-document summarization corpus manually annotated in European Portuguese. It summarises news articles from three areas, generalist, finance and sports newspapers.

Summ-it [21] : Brazilian Portuguese abstractive summarization corpus of "fifty texts from Science domain extracted from Science section of Brazilian daily newspaper Folha de São Paulo (FSP)" [21]. The summaries were produced manually.

CSTNews-Update [73] : This dataset is an update of an older dataset with the same name, the CSTNews [14]. This Brazilian Portuguese corpus provides summaries produced manually by annotators of news from Brazilian newspapers covering news of several areas. The multi-document capabilities were achieved by aggregating five clusters of similar news.

RulingBR [124] : This Brazilian Portuguese dataset includes several legal decisions from the Brazilian supreme court. This court is legally mandated to provide a summary of each decision produced. RulingBR is a single document abstractive dataset available for download in GitHub[8].

A Corpus for Sentence Compression [73] : The same research team that introduced CSTNews dataset [14], described above, also introduces this Brazilian Portuguese dataset. It analyzed one million news entries of the Brazilian digital newspaper G1 searching for news whose title and first sentence were similar. The title was then considered a compression of the first sentence.

Portuguese/Spanish Corpus for the Multi-Sentence Compression Task [87] : This dataset focused on sentence compression, a specific type of Text Summarization. It includes titles and summaries from Google News in Brazilian Portuguese.

WikiLingua: A Multilingual Abstractive Summarization Dataset [33] : This dataset includes wikiHow pages[9] and uses its abstract as a summary. The dataset is meant for single document abstractive text summarisation model training. The corpus is in Brazilian Portuguese.

XL-Sum [40] : A corpus based on the BBC news website is available in forty-four languages, including Portuguese. It is a single document corpus that uses the news body as a corpus and its lead as a summary. This corpus is in Brazilian Portuguese.

BrWac2Wiki [76] : A multi-document text summarization dataset that links the content of several Brazilian Portuguese webpages about the same topic to the correspondent abstract provided on that topic in the Portuguese Wikipedia. The Wikipedia abstract serves as a summary.

OpiSums-PT [23] : This corpus introduced both extractive and abstractive summaries for the same corpora. The dataset introduces summaries from books and reviews of electronic products. It includes materials exclusively in Brazilian Portuguese.

During our research, motivated by RulingBr, we identified the database from Portuguese legal decisions, DGSI[10], as a potential source of summarisation data. Portuguese law demand judges to include a structured summary of every legal decision published on this platform. However, we could not identify any dataset that compiles this information in a structured form.

[8] https://github.com/diego-feijo/rulingbr
[9] https://pt.wikihow.com/
[10] http://www.dgsi.pt/

3.7 Extractive Text Summarization

In the previous section, we introduced the generic topics that span extractive and abstractive summarization processes. In this section, we introduced the results for Portuguese ETS.

3.7.1 Training Process

ETS is widely used to solve summarization problems due to its simplicity compared to ATS models [32]. Abstractive Summarization is deeply connected with the NLP field because of the text generation requirements of these models. In contrast, ETS is more connected with the Information Retrieval field. The large majority of the extractive summarization techniques do not require NLP or ML knowledge [32].

For this reason, ETS architectures usually do not have a training step. Instead, statistical analysis like TF-IDF are used to determine the most relevant sentences in the corpus to include them in the final summary. However, motivated by the advances in NLP, ETS systems start to include neural methods to determine the most relevant textual parts to include in the summary.

3.7.2 Models

We were able to identify three ETS models for the Portuguese case. The results are summarized in Table 3.16 and described below.

Table 3.16: Analysis of Portuguese extractive summarization Tools ROUGE-1 performances. Including information regarding the training set (T.D), if it is a Multi-Document (M.D) dataset, the Portuguese Variant (Variant) of the corpus, and the Test Set.

Project Name	Authors	Year	T.D	M.D	Test Set	ROUGE-1
A Language Independent Algorithm for Single and Multiple Document Summarization [66]	Rada Mihalcea	2005	n/a	Y	TeMário	35,73
SIMBA: An Extractive Multi-document Summarization System for Portuguese [108]	Sara Silveira	2012	CSTNews	Y	CSTNews	45,57
RSumm [95]	Rafael Ribaldo	2012	n/a	Y	-	n/a

A Language Independent Algorithm for Single, and Multiple Document Summarization [66] : The authors propose an Iterative Graph-based algorithm to solve Brazilian Portuguese and English ETS. The authors are motivated by the results obtained by the PageRank algorithm [80] at Google.

SIMBA [108] : An Extractive multi-document model that uses TF-IDF combined heuristics to filter non-relevant content based on sentence length. The extractive capabilities of SIMBA were designed to operate in European Portuguese.

RSumm [95] : The authors propose a graph-based architecture for ETS. The proposal suggests abstracting the input text as a graph and performing the summary based on a graph traversal. This research was conducted for the Brazilian Portuguese case.

3.7.3 Results

The benchmarks for Portuguese ETS are presented in Table 3.16. We can observe that these models perform poorly, opening space for further research and the application of other summarization techniques.

3.8 Abstractive Text Summarization

In this section, we describe the training process, the models and the benchmark results of ATS models for the Portuguese case. As previously mentioned, ATS is a more challenging task than ES due to the subjectivity of quality judgement of the summary produced and the necessity for advanced text generation capabilities [32, 60]. However, new Large Language Models can surpass these challenges, which increased the attention on ATS models.

3.8.1 Training Process

Modern ATS systems based on Transformer architectures are typically unsupervised sequence-to-sequence architectures [60]. The goal is to train a model capable of generating a suitable summary for an input prompt.

This task's sequence-to-sequence nature derives from the fact that the next summary word is based on the previous sentences generated.

3.8.2 Models

We identified three projects that address ATS for the Portuguese case using modern transformers techniques. The results of this survey are compiled in Table 3.17 and described below.

Table 3.17: Analysis of Portuguese Abstractive Summarization Tools ROUGE-1 performances. Including information regarding the training set (T.D), if it is a Multi-Document (M.D) dataset, the Portuguese Variant (Variant) of the corpus, and the Test Set. Some of the test sets are defined by the authors(A.D).

Project Name	Authors	Year	T.D	M.D	Test Set	ROUGE-L
PLSUM [76]	Andre Seidel Oliveira	2021	BRWac2Wiki	N	BRWac2Wiki	27.10
Improving abstractive summarization of legal rulings through textual entailment [34]	Diego de Vargas Feijo	2021	RulingBr	Y	RulingBr	35.01
Deep Learning-Based Abstractive Summarization for Brazilian Portuguese Texts [81]	**Pedro H. Paiola**	2022	A.D	N	A.D	29.84

Deep Learning-Based Abstractive Summarization for Brazilian Portuguese Texts [81] : The authors fine-tuned the PT-T5 [15] model in a combination of several datasets–WikiLingua [33], XL-Sum [40], TeMário [84] and CSTNews [14]

Improving abstractive summarization of legal rulings through textual entailment [34] : This project applied a BERT architecture [28] to summarise legal documents using the dataset RulingBR [124].

PLSUM [76] : The authors used a Pre-Trained T5 for the Portuguese case, the PTT5 [16], to generate Wikipedia-like summaries to an input prompt.

3.8.3 Results

Table 3.17 presents the benchmarks obtained for ATS summarisation. We can observe that despite the low number of resources identified for both ES and ATS, the results that apply modern transformers in the context of ATS tend to perform better than older ES architectures. These results are promising, showing space for further improvements using Transformers architectures. More results must be catalogued to validate this tendency.

3.8.4 Research Teams

During our research, we identified some research teams worth mentioning in the field of Text Summarization. The first team we must refer to is the efforts of the Portuguese company Priberam, currently known in Portugal for its online dictionary, Priberam has published publicly the unique dataset in European Portuguese we were able to find, the Priberam Compressive Summarization Corpus [5]. The second project we want to mention is Sucinto, led by Thiago A. S. Pardo, at the University of São Carlos, São Paulo, Brazil. This team is associated with the introduction of

CSTNews [14] and other resources for multi-document summarization in the Brazilian Portuguese case. Sucinto is a significant source of summarization resources for the Portuguese case.

3.9 Relation Extraction

This section introduces the resources identified while researching RE. This task was already covered in Section 2.2.8. First, we briefly explain the training process for this NLP task. Then we list Portuguese datasets to support this task and the models trained to address it. We conclude by introducing the relevant stakeholders that present work in Portuguese RE.

3.9.1 Training Process

RE can be observed as supervised or semi-supervised learning [52]. The training goal is the same independently of the technique used: Train a classifier capable of outputting a suitable relation between NER tags. This fact forces RE models to operate with the output provided by the NER model.

3.9.2 Datasets

We were able to identify four RE datasets for the Portuguese case. The results are summarized in Table 3.18 and described below.

Table 3.18: Analysis of Portuguese RE Dataset properties.

Project Name	Author	Year	Corpus Type	Variant
ReRelEm [38]	Cláudia Freitas	2008	HAREM Adaptation	BOTH
PAPEL [77]	Hugo Gonçalo Oliveira	2010	Dicionário Porto Editora	PT-PT
CARTÃO [78]	Hugo Gonçalo Oliveira	2011	Dicionários Porto Editora	PT-PT
Summ-it++[36]	A. Antonitsch	2016	Newspaper	PT-BR
IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks[22]	Sandra Collovini	2019	HAREM Adaptation	BOTH

ReRelEM [38] : The authors added RE features to HAREM [69]. This is a similar work to the one conducted in the context of the IberLEF 2019 conference described below.

PAPEL [77] : This dataset is not a standard RE resource and only catalogues semantic relations between words in the corpus. However, it can be categorized as an RE resource since several RE

frameworks use rule-based systems based on semantic relation to delivering RE capabilities [52]. For this reason, we decided to mention this European Portuguese dataset in this section.

CARTÃO [77] : This dataset is an extension of the work developed in PAPEL. It is not a standard RE dataset, yet it was considered because the literature shows how semantic relations can perform RE [52].Section 2.2 briefly describes the eight NLP tasks we propose to catalogue the existing resources in Portuguese.

Summ-it++ [36] : This Brazilian Portuguese dataset extends Summ-it [21], a corpus dedicated to Text Summarization. Summ-it++ introduces manually labelled relations.

IberLEF 2019 conference [22] : Adaptation of HAREM [69] with RE features. HAREM is a NER dataset that compiles both corpora from European and Brazilian sources.

Motivated by the work introduced by David. S. Batista [8] that used Portuguese Wiki and DBPedia information to establish relations between entities, we decided to mention these two platforms as a source of RE corpora yet to be properly compiled.

3.9.3 Models

We identified a single model to address RE in Portuguese, the work presented by David S. Batista [8]. The author used the resources available by the Portuguese DBPedia and Wikipedia to determine the closest relations between different entities using a KNN algorithm associated with a Jaccard coefficient to determine existent relations.

Table 3.19: Analysis of Portuguese RE Models F1-Score Performance (F1-Score). Some of the test sets are defined by the authors (A.D)

Project Name	**Author**	**Year**	**Test Set**	**F1-Score**
Semi-Supervised Bootstrapping of Relationship Extractors with Distributional Semantic [8]	David S. Batista	2015	A.D	0.556

3.9.4 Results

The results obtained using Wikipedia and DBpedia score 55.6% F1-Score. There is space for further improvements in Portuguese RE systems. More contributions to this NLP task are required.

3.10 Semantic Role Labeling

In this section, we describe the literature review to assess the status of SRL resources in Portuguese. We briefly introduced this NLP task in Section 2.2.5. Now we focus on listing the SOTA training process, datasets and models for Portuguese SRL.

3.10.1 Training Process

SRL is a supervised NLP task. It is a sequence-to-sequence task since models aim to correctly label each token's argumentative role in the corpus to its semantic role [53].

3.10.2 Datasets

The SRL task introduces relevant information to discriminate the meaning of an input prompt, yet only a few datasets are available in English or Portuguese. We were able to identify three contributions to Portuguese SRL. The results are briefly described in Table 3.20 and described below.

Table 3.20: Analysis of Portuguese SRL datasets properties.

Project Name	Author	Year	Variant
Propbank-PT [9]	António Branco	2012	PT-PT
Propbank-Br [30]	Magali Sanches Duran	2012	PT-BR
Post-Scriptum Corpus [123]	Post-Scriptum	2020	PT-PT

Propbank PT [9] : Similar to Propbank-Br [30], the PT version of Propbank is the result of the efforts made at Lisbon Faculty of Science to create a European Portuguese version of the original Propbank [54]. This dataset was manually annotated, resulting in a high-quality data source for European Portuguese SRL.

Propbank Br [30] : A manually translated Brazilian Portuguese version of the widely used SRL English Dataset Propbank [54].

PS corpus (Post-Scriptum) - treebank [123] : This treebank offers SRL annotations of historical Portuguese and Spanish documents that span from the XVI to the XIX centuries.

3.10.3 Models

We identified two projects that introduce new models for SRL Portuguese. The results are summarized in Table 3.21 and described below.

LX-SRLabeler : This production of the NLX group at Lisbon Faculty of Science needs to be better documented. No paper describing it is available, yet a short description on the tool's website states that it "uses probabilistic grammars" and obtained 70% F-Score.

Semantic Role Labeling in Portuguese [75, 79] : In this project, Sofia Oliveira used the PropBank.Br to train an architecture composed of a BERT base encoder combined with a decoder based on the Viterbi algorithm.

Table 3.21: Analysis of Portuguese SRL Models F1-Score Performances detailing the European (F1-PT) and Brazilian Portuguese (F1-BR) F1-Score performance

Project Name	**Author**	**Year**	**Methodology**	**Test Set**	**F1-PT**	**F1-BR**
LX-SRLabeler[11]	António Branco	2014	Probabilistic Grammars	n/a	0.700	-
Semantic Role Labeling in Portuguese [75, 79]	Sofia Oliveira	2021	BERT and Viterbi Decoder	PropBank.Br	-	0.782

3.10.4 Results

The two projects obtained similar results by applying different methodologies. The results are above 70% F1-Score, a good result for a challenging task like SRL. However, the sample is scarce, and more investigation needs to be performed to validate the results presented.

Chapter 4

Building Portuguese Natural Language Resources

In the previous chapter, we extensively analyse Portuguese resources to support eight NLP tasks. In this chapter, we clarify the problem statement (4.1), present the main hypothesis that orients this document (4.2) and formulate the four research questions we propose to answer in the conclusion stage of our project (4.3). In Section 4.4, we explain the selection of NER and ATS as the object of implementation in practice of the methodology followed (4.5). We conclude this chapter by listing software architectural decisions relevant to understand the following chapters of this document (4.6).

4.1 Problem Statement

The analysis of Portuguese Natural Processing SOTA (3) make clear the following problems:

1. The resources are insufficient
2. Several resources are not ready to work off-the-shelf
3. The resources are dispersed across different platforms and formats
4. The techniques available in several NLP tasks are not SOTA

The insufficient number and extent of resources is a common critique pointed out by most of the researchers we catalogued in the SOTA revision chapter of this document. Authors point to inadequate datasets as a significant constraint in their research process and the results obtained.

Our literature review showed that Portuguese NLP Resources do not follow the most recent standardization trend of formats. Each research team publishes their resources in their way, without a particular concern of unification with its peers. By no means does the majority of resources offer simple off-the-shelf utilization. Portuguese NLP lacks an official platform specialized in the indexation of these resources. In the absence of a dedicated offer, we were able to identify some

personal contributions, like David S. Baptista's GitHub[1] and Portulan Clarin[2]. However, these platforms serve the creator's interests and do not provide a macro coverage of the NLP field. For example, David Baptista is concerned with NER and RE; therefore, it only provides NER and RE resources to the public. Portulan Clarin, on the other hand, is managed by the NLX group at the University of Lisbon and focuses mainly on cataloguing resources produced at that university.

Furthermore, our survey demonstrates that the SOTA results for several NLP tasks are already several years old. Only some cases identified apply modern Transformer-based SOTA techniques. The ones that applied it tended to beat new SOTA benchmarks.

4.2 Main Hypothesis

We intend to develop a new methodology to extend Portuguese NLP resources that we summarise in the following hypothesis:

> *In a low-resource language like Portuguese, combining diverse sources and including synthetic data to generate a larger training corpus will allow pre-trained transformer models to approach the benchmarks obtained in English, bridging the gap between the two languages.*

This hypothesis derives from the conclusions extracted from our literature review. We observe that transformer-based solutions outperform the scores in all Portuguese NLP tasks where they were introduced. Also, there are some projects where combining datasets was possible but not tested, opening new research possibilities.

4.3 Research Questions

This section presents the four research questions (RQ) we considered paramount to guide our work plan:

RQ1: Does combining datasets with the same labelling scheme but different sources positively impact the overall performance of models?

RQ2: What is the additional overhead to introduce silver labelled data in the training pipeline?

RQ3: Does European Portuguese data negatively impact Pretrained Brazilian Portuguese transformer models?

RQ4: Which are the best public framework to deploy an off-the-shelf NLP tool?

[1] https://github.com/davidsbatista
[2] https://portulanclarin.net

4.4 NLP Tasks Selected

We decided to demonstrate our methodology applied to Named Entity Recognition (5) and Abstractive Text Summarization (6). We selected NER due to its relevance in NLP. NER is used by many high-level NLP tasks to improve their results. Furthermore, for the Portuguese case, the usage of NER is dominated by the off-the-shelf models provided by spaCy. The three Portuguese NER models this Python library provides offer an easy-to-use but with poor performance solution. We intend to provide an alternative to spaCy by introducing not only an easy-to-use, off-the-shelf solution, but also a significant increase in the performance levels for the Portuguese case.

Abstractive Text Summarization was selected since it is one of the NLP tasks for which we identified more resources during our survey. This fact makes it a natural decision to test how our central hypothesis based on combining corpus from different sources performs.

4.5 Methodology

We establish an incremental protocol to fulfil our goal of creating SOTA off-the-shelf solutions for NER and ATS. Our goal is to increase the complexity of the techniques applied linearly. Simpler approaches to the problem are tested first; if these techniques reveal insufficient, we move to more challenging methods. We list the guidelines of this protocol in the following subsections.

4.5.1 Use Pre-Trained Transformer Architectures as Baseline

We intend to use the pre-trained Transformer architecture that performs better in a particular NLP task as starting ground for our research. We take advantage of the information we collected during our literature review to determine BERT-CRF [115] and Portuguese T5 [81] as the promising architectures for NER and ATS, respectively. We expect that applying our methodology to these architectures outperforms the current benchmarks in these two NLP tasks.

4.5.2 Dataset Definition

Defining a dataset is an essential step in an ML problem. Our main hypothesis (4.2) states that this corpus must be obtained in two stages. First, we experiment how merging datasets from different sources that share the same labelling scheme impact the overall models' results. Then, if the results do not improve the current SOTA benchmark, we intend to apply DA techniques to achieve this goal. We visually represent this proposal in the following sequence diagram (4.1).

4.5.3 Data Augmentation

Data Augmentation was the first NLP task introduced in this document (2.2.1) due to its relevance in extending corpora in low-resource languages. However, due to several factors, mainly lack of time, it is not feasible to explore this complex NLP field to a full extent.

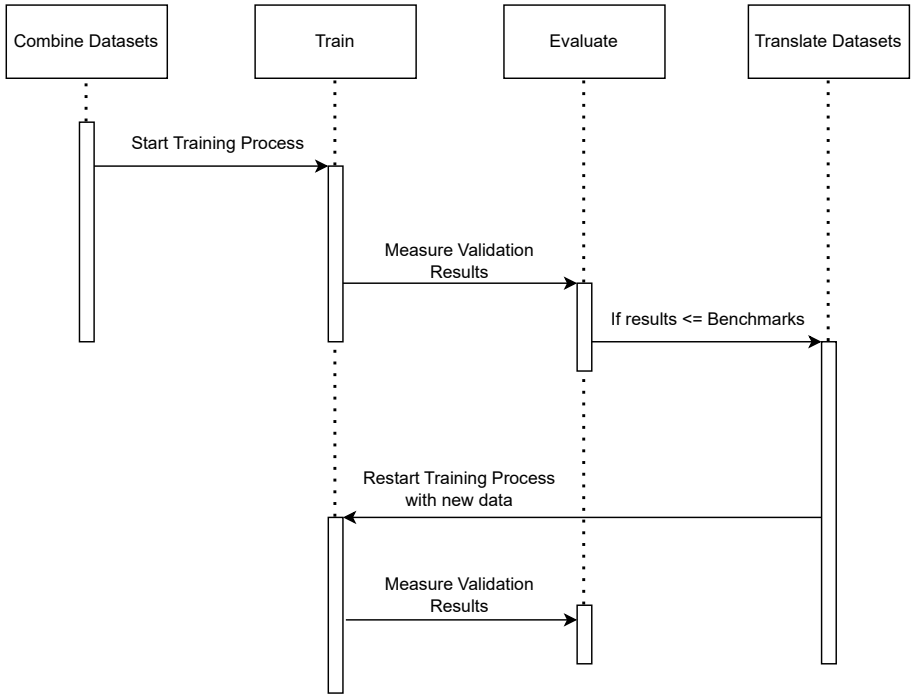


Figure 4.1: The chronology of the methodology proposed.

We intend to restrain our attention to DA based on Machine Translation. We apply MT as an escape method if our central hypothesis based on dataset combination reveals insufficient to outscore the current benchmarks. Formally, the DA transformation we intend to apply is the Machine Translation process. This approach requires an extra label alignment step to preserve the order of the labels in the corpus (2.2.1).

4.5.4 Evaluation

Machine Learning research requires an evaluation step to measure the performance of the models produced. In order to improve comparability with other NLP researchers, we intend to use standard evaluation metrics in both NLP tasks. For NER, we focus on quantifying the F1-Score of our models, while in ATS, we use ROUGE-L in the evaluation stage.

To ensure correctness and sound ML practices, all the validation performed will ensure the test set is part of a gold-labelled dataset provided by an external party, enforcing transparency and confidence in the results obtained.

4.6 Architectural Decisions

This section introduces the major software architectural decisions we took in this dissertation. First, we introduce the pipeline design pattern, the structural pattern that modules all the code

produced. Then in Subsection 4.6.2, we explain our hybrid approach that combines using Huggingface with Pytorch code to solve the problem. We conclude by introducing the three third-party components used in this dissertation, Deep-Translator (4.6.3), PaperGradient Space[3] and GitHub[4].

4.6.1 The Pipeline Design Pattern

The idea of task sequencing is a feature provided in all significant ML frameworks due to the sequential nature of ML problems [105]. Organising the code following a pipeline philosophy simplifies the understanding, ensures encapsulation and avoids undefined behaviours in the code. This way, we lower the chance of bugs that could affect the confidence in the results obtained. To enforce this design pattern, we defined a Python Object Oriented construction of it, and only when necessary, we introduced an external Makefile to ensure the sequential behaviour of our code.

4.6.2 Hybrid Architecture: Combining HuggingFace with Pytorch

Motivated by the idea of exploring pre-trained solutions to deliver off-the-shelf models, we considered that the platform that currently suits this intent better is HuggingFace. The abstractions provided by HuggingFace allow faster deployments and easiness of use but come at the cost of restraining programmers from executing low-level tensor operations. In order to overcome this situation, we developed a hybrid approach that integrates Pytorch code with Huggingface solutions. This approach simplifies the public deployment and compatibility with a wide range of datasets and models available in the HuggingFace platform.

4.6.3 Translation Library: Deep-translator

Deep translator is a primary tool to obtain access to translation services embedded within Python code [26]. The capability of embedding translation features in pipelines increases the flexibility and automation of the solutions presented. Deep-Translator supports eleven translation engines, some of which, like Google Translator, are free, but it is also possible to integrate with paid services like DeepL and MS Azure. Free services typically implement limits in the rate of daily requests; the paid services require an API key as an input argument to the Deep-Translator library.

4.6.4 GPU Training: Papergradient Space

Training modern Transformer based models requires access to GPUs. We selected the PRO tier of Paperspace Gradient from the available alternatives to serve this purpose. This service provides access to six hours of continuous training at a monthly cost of $8. This service offers a Jupyterlab interface to modern NVIDIA GPUs specialised in ML training. Our pipelines are tested in the following hardware: NVIDIA P5000, RTX5000 and A4000.

[3] https://www.paperspace.com/gradient
[4] https://github.com/

4.6.5 GitHub

One of the goals of PT-Pump-Up, the project that motivates this dissertation, is to create a Hub for Portuguese NLP resources. The content researched during this dissertation is the first batch of information published on this platform. In order to ensure accessibility and explore automation capabilities, we decided to use GitHub as the tool to achieve this goal more efficiently.

Chapter 5

Portuguese Named Entity Recognition

This chapter introduces the results obtained during our research on Portuguese Named Entity Recognition. We first introduce the labelling schemes we decided to work with (5.1). Then in Section 5.2, we introduced the current SOTA architecture that served as baseline during our research on Portuguese NER. We present the results obtained in Sections 5.3 and 5.4. Finally, we conclude this chapter by introducing a comparative performance analysis in the same evaluation conditions of our models with the Portuguese NER solutions provided by spaCy (5.6).

5.1 Dataset Definition

Our survey of the current SOTA of Portuguese NER (3.3) concluded that Mini HAREM is the primary source of benchmarking extraction for this NLP task [69]. Furthermore, we concluded that the English dataset CoNLL-2003 is the most used in this field [120]. The importance of CoNLL-2003 for this NLP task turned its labelling scheme into an unofficial standard that was followed by many other datasets in recent years (2.2.4).

The HAREM dataset has two labelling schemes, the HAREM default and the HAREM selective (a subset of the Default case). Both cases do not follow the CoNLL-2003 standard. Due to these differences and the relevance of all three formats, we divided our task into three small substeps. We propose to apply the same ML pipeline to the three labelling schemes; this proposal motivates us to create a flexible, well-constructed ML pipeline capable of handling different labelling schemes without triplicating parts of the code. We detail the specificities of the three labelling schemes in Table 5.1).

5.1.1 Time Distribution Among Labeling Schemes

HAREM Default presents a very detailed labelling scheme, while CoNLL-2003 presents a considerably simplified approach to the NER problem; CoNLL-2003 simplifies the challenge of NER by aggregating much of the information in its MISC category. Motivated by the reasons above, we focus our attention mainly on Harem Selective. The remaining time was split across the other two

Table 5.1: Labeling scheme of the NER datasets considered. Including information about the regularization of the dataset in the BIO format(BIO), Number of entities considered(N.E)

Dataset Name	**BIO**	**N.E**	**Entities**
HAREM Default	N	10	Person, Organization, Location, Value, Date, Title, Thing, Event, Abstraction, and Other
HAREM Selective	N	5	Person, Organization, Location, Value, and Date
CoNLL-2003	Y	4	Person, Organization, Location, Miscellaneous

sub-tasks. The simplification provided by CoNLL-2003 demanded less time than Harem Default to achieve the desirable results.

5.2 BERT-CRF

This section describes the Transformed based architecture used during our research in Portuguese NER. The results obtained by Souza et al. [115] while using the BERT-CRF are promising for Portuguese NER. Motivated by those findings, we developed a pipeline capable of training this architecture in the HAREM labelling scheme and any labeling scheme supported by HuggingFace. We begin this section by detailing this architecture (5.2.1). Then, we present some benchmarks obtained by BERT-CRF in other languages (5.2.2). We conclude by enumerating the contributions from a software engineering point of view that our BERT-CRF pipeline introduces when compared with the preexistent works in this topic (5.2.3).

5.2.1 Architecture

The starting point of Portuguese BERT-CRF is the usage of Portuguese BERT, BERTimbau, in its pre-trained format [114]. The idea is to benefit from the knowledge captured by this LLM during its unsupervised learning step to boost the results for Portuguese NER. A CRF layer is also introduced after the LLM. We identified two options to transform the BERT output logits into NER labels. The first is to use a softmax operator, and the second is to delegate the token labelling stage to a CRF layer [45]. Hu et al. state that "Although the softmax function outputs the label corresponding to the maximum probability of the word, the output label is independent of each other, which means that it has a weak relationship with the context, resulting in a decrease in accuracy" [45]. This motivated us to select the CRF approach.

The BERT architecture accepts input sizes between 0 and 512 tokens (BERT_MAX_SIZE), and the typical output size (BERT_HIDDEN_SIZE) is 768. The CRF layer requires an input size equal to the original sentence. Therefore, a size normalization step was necessary to turn the BERT output layer to the same size as the input sequence. These requirements are evident in the visual representation provided (5.1).

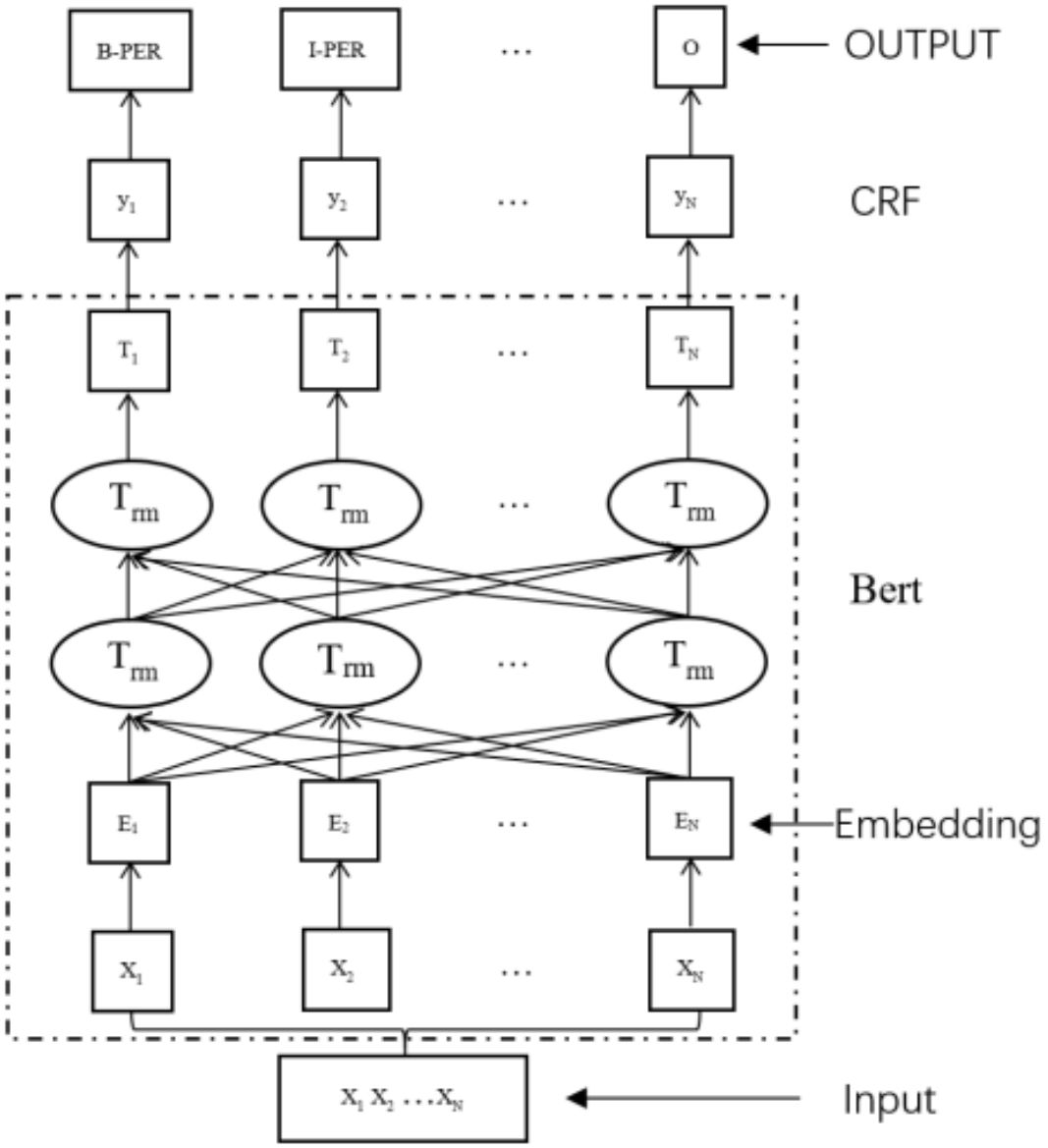


Figure 5.1: BERT-CRF Architecture [45]

To enforce the correct tensor size, we stacked a simple linear feed-forward layer between the LLM output and the CRF input to perform this reduction. Nevertheless, due to architecture limitations, an additional mechanism based on padding is necessary to handle the input sequence size variability. All sequences are padded to the same size; this extra padding is signalled to the CRF layer with an extra token mask tensor which denotes which tokens should be ignored by this model.

To sum up, we use BERTimbau to take advantage of the knowledge of the Portuguese language this LLM encapsulates; a linear layer is needed to perform a sequence size normalization step, and the final stage includes a CRF model to perform the sequence labelling stage to take advantage as much as possible of the relation between tokens.

5.2.2 Results Obtained

As mentioned, we decided to use this architecture because it represents the more reliable SOTA benchmarking architecture for Portuguese NER. Moreover, this architecture was successfully applied in NER pipelines dedicated to other languages. The Table 5.2 presents some NER benchmarks for those languages.

Table 5.2: F1-Scores of NER BERT-CRF architectures in several different languages. Presenting the number of documents (N.C) that compose the dataset

Project Name	**Author**	**Year**	**Test Set**	**N.C**	**F1-Score**
Portuguese Named Entity Recognition using BERT-CRF [113]	Fabio Souza	2018	HAREM	257	0.832
BERT+CRF based Named Entity Recognition model for Korean[1]	Joosung Yoon	2021	ModuNEKorpus	20,188	0.876
Chinese Named Entity Recognition based on BERT-CRF Model [45]	Shulin Hu	2022	People's Daily	27,818	0.945

These results demonstrate that this architecture scores well with little effort for different languages. The Portuguese case is from the three cases presented as the one that performs worse. It is possible to observe a correlation between the F1-Score obtained and the dimensionality of the dataset used to train and test.

5.2.3 Improving Code Accessibility

Even though our architecture is based on previous literature, we extend the work of existing BERT-CRF solutions by delivering a flexible pipeline with clear gains from a Software Engineering perspective compared to its counterparts. First, it allows training in every type of Token Classification task, not only NER. Secondly, because it is agnostic to the datasets used, it is not focused on any particular case. Moreover, our solution allows controlling relevant training parameters using a high-level configuration file without modifying the source code. Additionally, our solution offers automatic deployments to HuggingFace, simplifying the process of making these projects publicly available in an off-the-shelf manner. Table 5.3 compares our pipeline and the current SOTA for the Portuguese language focusing on the features implemented in both proposals.

Table 5.3: Comparative analysis of the features implemented between Current SOTA Project (Current SOTA) and Our Proposal.

Feature	Current SOTA	Our Proposal
Load From HuggingFace	N	Y
Train HAREM	Y	Y
Train Other Datasets	N	Y
Train Other Token Classification Tasks (e.g., POS Tagging)	N	Y
Parameters Depend on Configuration File	N	Y
Integration with PyPlotly: Automatic Generation of Plots	N	Y
Automatic Deployments to HuggingFace	N	Y

5.3 Combining Different Portuguese NER Datasets

This section presents the results obtained while combining different NER datasets with similar labelling schemes into single units. We divide the results obtained during this stage into three subsections, one for each labelling scheme studied–HAREM Selective, HAREM Default and Conll-2003.

During our literature review on Portuguese NER, we observed that the training process of Portuguese BERT-CRF relied exclusively on the small dataset First HAREM for training and validation [113]. We saw an opportunity to improve the results in this reduced amount of resources used. Our intuition was that if we combined the little data identified in our literature review for Portuguese NER (3.7), we would obtain at least marginal gains compared with the baseline. This statement condenses the central hypothesis of this dissertation (4.2), a strategy with incremental complexity, which only applies complex tasks if simpler ones prove ineffective.

The process of combining datasets requires four stages. The first stage is loading the files from the source, either local or remote. The second stage requires parsing of the files; different datasets require different parsing codes; some are in XML, others in JSON, and within the same file format, they have different annotation schemes to be parsed. After the information is parsed, there is a label normalisation stage, where a hand-crafted set of rules normalise the source labelling scheme to match the destination format. This stage is critical; the results vary dramatically depending on the rules defined. The final stage includes the pushing of information to the Huggingface hub.

5.3.1 HAREM Selective Results

This subsection lists the results obtained while combining different data sources into single training unit following a HAREM Selective labelling scheme. In these experiments, the test set was

mini HAREM [69] since this is the most used benchmarking dataset for Portuguese NER. We started by defining a separate validation set to reduce the bias on the final scores obtained, yet, we quickly abandoned that possibility due to the decrease in F1-Score we faced as the result of splitting the training set to serve as validation. For this reason, the results listed (5.4) do not use the concept of validation dataset; we collect benchmarks directly on the test set.

Table 5.4: Results of combining Pre-Existent NER datasets with HAREM Selective. Describing all the training parameters, the Dataset (T.D), the BERT model used(BERT), the Learning Rate(L.R), the batch size(B.S), the Input Sequence Length(Seq Len), if any BERT Layer was freezed(L.F), the F1-Score and the variation of the F1-Score obtained to current SOTA (V.S)

T.D	**BERT**	**L.R**	**B.S**	**Seq Len**	**L.F**	**F1-Score**
F. HAREM	Base	1e-5	32	128	N	0.802
F. HAREM	Large	1e-5	16	128	N	0.832
Portuguese MAPA	Large	1e-5	16	128	N	0.149
F. HAREM + S. HAREM	Base	1e-5	16	400	N	0.723
F. HAREM	Base	3e-5	16	400	6	0.794
F. HAREM	Base	3e-5	16	400	9	0.776
F. HAREM	Base	3e-5	16	400	11	0.700
Two Stage Training First: F. HAREM Second: S. HAREM	Base	First: 1e-5 Second: 1e-6	16	400	N	0.806
S. HAREM	Base	1e-5	16	400	N	0.730
F. HAREM + Ontonotes 5.0-PT	Large	1e-5	16	128	N	0.790

As stated at the beginning of this chapter, HAREM selective was the first labelling scheme to be studied and the one we spent more time developing. For this reason, it is essential to understand the following conclusions extracted from the HAREM selective results (5.4) since they influence the experiments for the remaining two NER labelling schemes.

The nine experiments listed tried different approaches to achieve the ultimate goal of elevating the SOTA benchmarks for Portuguese NER on HAREM Selective. Unfortunately, we were only able to match the current benchmark result. BERT-Large performs slightly better than BERT-Base. The best results were obtained for learning rates in the magnitude of e-5. Our attempts to improve results by increasing the embedding input size, which results in a more contextualized input, did not reveal paramount for the final result, but it introduced more GPU RAM demands. We also varied the batch size but did not reveal any impact. We explored freezing the layers of BERT; the

results clearly show that this technique lowers the overall results obtained and, therefore, should be avoided. In Table 5.5, we summarise the optimal parameters that maximize F1-Score and have lower hardware demand for HAREM Selective. These findings serve as starting point for the remaining two experiments on NER.

Table 5.5: Optimal parameters for HAREM Selective training

Parameter	**Value**
BERT-Model	Large
Learning Rate	1e-5
Batch Size	16
Sequence Length	128
Number Freezed Layers	0

Finally, we conclude that introducing other datasets tends to include entropy in the final result. This fact is evident in the performance decrease observed when Second HAREM is introduced in the training set; even though these two datasets share the same LS and annotation team, it evidences that dataset combination is a more complex process than expected. These findings compromise our initial intuition that a simple dataset combination would elevate the training results by introducing a bigger input.

5.3.2 Combining Second HAREM and MAPA datasets with HAREM Selective

In this series of experiments using pre-existent NER datasets, we introduced Second HAREM and Portuguese MAPA in addition to the First HAREM and MiniHAREM. The second HAREM follows the First HAREM labelling scheme, whereas the MAPA dataset requires a manual step of labelling normalization. Table (5.6) presents the conversion map designed to converge MAPA into HAREM Selective format.

Table 5.6: MAPA To HAREM Selective conversion map

BIO-Tag MAPA	**BIO-Tag Harem Selective**
O	O
BUILDING	LOC
CITY	LOC
COUNTRY	LOC
PLACE	LOC
TERRITORY	LOC
UNIT	VALUE
VALUE	VALUE
YEAR	TIME
[TIME] STANDARD ABBREVIATION	TIME
MONTH	TIME
DAY	TIME
AGE	VALUE
ETHNIC CATEGORY	O
FAMILY NAME	PER
INITIAL NAME	PER
MARITAL STATUS	O
PROFESSION	O
ROLE	PER
NATIONALITY	O
TITLE	PER
URL	LOC
ORG	ORG
VEHICLE	O
TIME	TIME

5.3.3 HAREM Default Results

This subsection lists the results obtained while combining preexisting datasets in the HAREM Default labelling scheme. For testing, we used the preexisting Mini-HAREM in the same labelling scheme. This subsection already implements the primary takeaways extracted from our research on HAREM Selective (5.3.1). As previously stated, the definition of a third validation set was initially tested, but the results were negatively impacted. Therefore, the results listed in Table 5.7 are tested directly in the test set.

Table 5.7: Results of combining Pre-Existent NER datasets with HAREM Default. Describing all the training parameters, the Dataset (T.D), the BERT model used(BERT), the Learning Rate(L.R), the batch size(B.S), the Input Sequence Length(Seq Len), the F1-Score and the variation of the F1-Score obtained to current SOTA (V.S)

T.D	BERT	L.R	B.S	Seq. Len.	F1-Score
F. HAREM	Base	1e-5	32	128	0.749
F. HAREM	Large	1e-5	16	128	0.767
F. HAREM + S. Harem 5.0-PT	Large	1e-5	16	128	0.724
F. HAREM + Ontonotes 5.0-PT	Large	1e-5	16	128	0.710
S. HAREM	Large	1e-5	16	128	0.693

We performed five different experiments for the HAREM Default labelling scheme. As expected, the additional complexity introduced by this labelling scheme, impact negatively the F1-Scores. The results obtained while training with Second HAREM reveal a pattern that spans from the HAREM Selective case. Even though these two datasets follow the same annotation scheme, they present differences that result in the introduction of entropy in the training process. Unfortunately, surpassing the current SOTA for this labelling scheme using mechanisms based on corpus combination was not possible.

5.3.4 CoNLL-2003 Results

This subsection lists the results of combining datasets for the CoNLL-2003 format. Testing this sub-task introduced two significant changes from the previous two. Firstly, there is no benchmarking dataset used for Portuguese CoNLL-2003, contrary to the other two cases where the literature uses Mini Harem as test set. Secondly, while training with the Portuguese subset of Wikineural [118], it was possible to include a third set of data to validate the train in each epoch. The authors of Wikineural predefined this validation set; we applied it without further adjustments. The results for the CoNLL-2003 labelling scheme are summarised in Table 5.8.

Table 5.8: Results of combining Pre-Existent NER datasets CoNLL-2003 format. Describing all the training parameters, the Training Dataset (Train.D), the Validation Dataset (V.D), the Test Dataset (Test.D), the BERT model used(BERT), the Learning Rate(L.R), the batch size(B.S), the Input Sequence Length(Seq Len) and the F1-Score

Train.D	**V.D**	**Test.D**	**BERT**	**L.R**	**B.S**	**Seq Len**	**F1-Score**
Wikineural	Wikineural	Wikineural	Large	5e-6	16	128	0.951
F. HAREM. CoNLL-2003 Format	-	Mini-HAREM. CoNLL-2003 Format	Large	1e-5	16	128	0.781
F. HAREM. CoNLL-2003 Format + S. HAREM. CoNLL-2003 Format	-	Mini-HAREM. CoNLL-2003 Format	Large	5e-6	16	128	0.756
Wikineural	Wikineural	Mini-HAREM. CoNLL-2003	Large	1e-5	16	128	0.535

As expected, under the same training parameters, this LS is the one that obtains the best F1-Scores. The reduced number of labels introduced by this task simplifies the process of NER. We establish a new benchmark for this architecture of 0.951 F1-Score in the Wikineural dataset. Again, two phenomena previously observed happened. First, the techniques based on dataset combination negatively impacted the overall F1-Scores obtained. Secondly, combining the First and Second HAREM introduces entropy in the final results.

5.3.5 Creating a CoNLL-2003 Version of HAREM

Unlike Wikineural, the HAREM dataset does not follow a CoNLL-2003 LS. To measure the effectiveness of this dataset in this LS, a normalisation step was required in the CoNLL-2003 format. The conversion map used is provided in Table 5.9.

Table 5.9: HAREM Selective to CoNLL-2003 conversion map

BIO-Tag HAREM Selective	**BIO-Tag CoNLL-2003**
O	O
PER	PER
ORG	ORG
LOC	LOC
TEMPO	MISC
VALOR	MISC

5.3.6 Conclusion

We perform three separate experiments within the context of dataset combination to measure the effectiveness of BERT-CRF for Portuguese NER. We focus our time mainly on HAREM Selective; for this reason, we have already extracted some conclusions for these particular cases. The incapacity to improve the benchmarks in this LS motivated some changes in the training process of the other two NER experiments. We could not improve the results in the HAREM default case (5.7). Fortunately, we achieved greater success in the CoNLL-2003 case by establishing a new benchmark for the Wikineural dataset with 0.951 F1-Score. Furthermore, we established the first benchmark for a CoNLL-2003 version of the HAREM dataset with 0.781 F1-Score. Motivated by these results, we present in the following section a new approach to the Portuguese NER problem based on data augmentation using Machine Translation (5.4).

5.4 Token Classification Translation Pipeline

This section describes the results obtained while exploring data augmentation techniques based on Machine Translation. We decided to study the application of this methodology after being unable to outperform the current benchmark in the HAREM default and the HAREM selective labelling scheme using dataset combination. Since we achieve greater success in the CoNLL-2003 case, we do not focus on augmenting the data for this particular case.

To support this task, we developed a multi-lingual translation pipeline based on the HuggingFace ecosystem capable of augmenting data for NER or any other token classification task. In the following subsections, we present a chronological analysis of the challenges this technique introduced and the solutions we found to address them.

5.4.1 Identify a Non-Portuguese Dataset Similar to HAREM: The Ontonotes 5.0

The first step to augment the HAREM dataset required a process of literature revision to determine a non-Portuguese dataset with a similar labelling scheme. A labelling scheme similar to HAREM is required to ensure the label normalization process is not lossy. This task revealed challenging since the Harem Default variant has ten NER labels. Our research identified Ontonotes 5.0/Conll-2012 [88] as the best match to those requirements. Table 5.10 presents a comparative analysis of the similarities between the labelling schemes of these datasets.

Table 5.10: Ontonotes 5.0 to HAREM Selective conversion map

Label	**BIO-Tag Ontonotes 5.0**	**BIO-Tag Harem Default**
Outside	O	O
Person	PER	PER
Nationalities or religious or political groups	NORP	ABSTRACAO
Facility	FAC	LOC
Organization	ORG	ORG
Geo-Political	GPE	LOC
Location	LOC	LOC
Product	PRODUCT	COISA
Date	DATE	TEMPO
Time	TIME	TEMPO
Percentage	PERCENT	VALOR
Money	MONEY	VALOR
Quantity	QUANTITY	VALOR
Ordinal	ORDINAL	VALOR
Cardinal	CARDINAL	VALOR
Event	EVENT	ACONTECIMENTO
Work of Art	WORK_OF_ART	OBRA
Law	LAW	Outro
Language	LANGUAGE	ABSTRACAO

In Table 5.10, we demonstrate that Ontonotes 5.0 has a one-to-one alignment to all the pre-existent Harem Default/Harem Selective labels.

5.4.2 Architecture

In order to establish a HAREM version of the Ontonotes 5.0 dataset, we were required to create a translation step. Our translation solution follows the pipeline design pattern (4.6.1) to ensure the easy maintainability and usage of the code. Our pipeline has three major stages. First, it translates the English dataset using the Python library Deep-Translator; Then it uses Awesome Align [29], a neural model specialized in cross-language token alignment based on the BERT-Multilingual architecture. The resulting alignments are parsed to support the reconstruction of the NER label sequence in Portuguese using the heuristics detailed in Subsection 5.4.3. Figure 5.2 presents a visual of the major steps of the pipeline.

5.4.3 Post-Alignment Step

The limitations enumerated in the Subsection 5.4.5 required us to enforce strict policies of Sequence Labelling Reconstruction. The reconstruction process is required since the output of the alignment step are pairs of indexes mapping tokens in the source to the destination language; we need to reconstruct the labels in the target language using the information presented in those alignments.

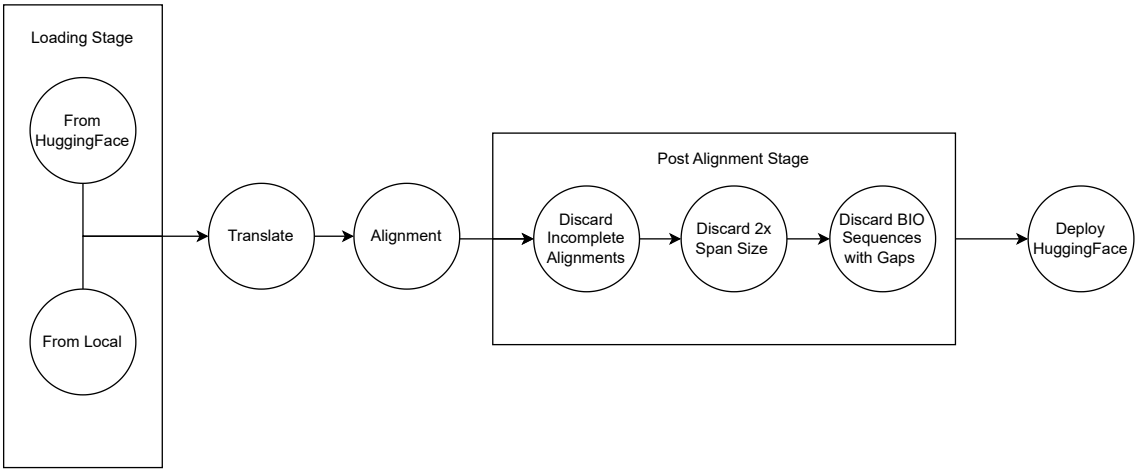


Figure 5.2: Alignment pipeline scheme.

In order to assess the numerous amount of incomplete outputs, we decided to discard all the faulty cases. Then, we also decided to limit the impact of model hallucination by discarding all cases whose span is twice the size of its length in the original language. Finally, we included an extra step of avoiding documents where a perfect sequence of B-Tags followed by I-Tags is not observed.

The consequence of these policies is that most of the original Ontonotes entries were discarded. We present those results summarised in Table 5.11. Despite the constraints, Ontonotes 5.0-PT contain documents with several times more information than the original HAREM.

Table 5.11: Variation in the Document Numbers(D.N) post-alignment. Comparative analysis Portuguese and English

Split	**English D.N**	**Ontonotes 5.0-PT D.N**	**Variation**
Train	10539	1900	-82%
Validation	1370	279	-80%
Test	1200	163	-87%

5.4.4 Results

This section presents the F1 scores obtained while training BERT-CRF models using silver labelled data from our pipeline of translation and alignment. Table 5.12 summarises the training results obtained using the Ontonotes 5.0 dataset to train a Named Entity Recogniser. The authors of Ontonotes 5.0 predefined the split in Train, Validation and Test set; we preserved them to achieve results in similar conditions.

Table 5.12: F1-Scores using Ontonotes 5.0 dataset. In bold, in the last row of this table, we introduce the current SOTA for English, Dice Loss for Data-imbalanced NLP Tasks [58] trained in original English Ontonotes.

Project Name	**Language**	**Train Set**	**Validation Set**	**Test Set**	**F1-Score**
Our Solution	Portuguese	Ontonotes 5.0-PT-Harem-Selective	Ontonotes 5.0-PT-Harem-Selective	Harem Selective	0.558
Our Solution	Portuguese	Ontonotes 5.0-PT	Ontonotes 5.0-PT	Ontonotes 5.0-PT	0.587
Dice Loss for Data-imbalanced NLP Tasks	**English**	**Ontonotes 5.0**	**Ontonotes 5.0**	**Ontonotes 5.0**	**0.920**

The results obtained from training with the translated version of Ontonotes 5.0 have a poor F1-Score. We performed two different experiments to validate these results. First, we train with Ontonotes 5.0-PT and test it with the Mini HAREM-Selective; the results have a difference in F1-Score around 40% to our best-performing model (5.4). Secondly, to ensure the downgrade observed was motivated by the corpora, not the pipeline, we trained and tested the BERT-CRF using translated data and scored marginal gains to the result obtained in HAREM. Furthermore, the difference between both results in the translated set and the original benchmark for the English language is significant. This gap demonstrates much work to be done to improve the Portuguese results.

5.4.5 Limitations

This mechanism of Data Augmentation based on MT presents several limitations that negatively impact the results obtained. We categorize this limitation into two orders. First, limitations at the software engineering level regarding the pipeline. Secondly, restrictions on the quality of the alignments obtained. We explore in more depth these two situations in the following paragraphs.

Current NLP datasets oriented for deep learning contain thousands or even millions of documents. Translating that amount of documents either costs vast sums of money in paid services like AWS or MS Azure or are hard constrained by the daily and per-second limits in the number of API requests imposed by free services like Google Translator. Due to budget constraints, we were forced to use Google's free services. Executing the translation job for Ontonotes 5.0 would take 49 hours without further engineering. To overcome this situation, we introduced multi-threading in our translation pipeline, which shortens the translation periods at the expense of additional synchronisation problems. Firstly, multi-threading quickly depletes the number of requests to the Google services allowed each second. To fix this situation, we were forced to reduce the number of threads concurring with each other and introduced some busy wait mechanisms to avoid excessive parallelism. Secondly, by introducing multi-threading, the order of the documents is lost; if this information is a requirement, other techniques must be included to ensure the correct sorting.

To sum up, we implemented a multi-threading translation pipeline with reduced threads, busy wait and file logging to overcome the challenges of concurrency. Our solution reduces the translation time of Ontonotes 5.0 to 40 minutes.

Lastly, we were surprised by the poor quality of the results obtained by off-the-shelf token alignment models. Previously in this document, we acknowledge that data augmentation in NER is challenging due to the need to preserve the correct ordering of tokens and labels in a sentence, which could easily be lost during a DA process. Some literature introduces methodologies to overcome this situation [47, 31, 29]. However, none of them produces satisfactory results for the English-Portuguese case. Our final release uses Awesome Alignments [29] not because we have evidence that it scores better than the other two cases but because it is based on HuggingFace, making it simpler to integrate with our pipeline.

5.4.6 Conclusion

The results demonstrate that our pipeline requires improvements to score SOTA results for Portuguese NLP and equalise the results obtained in similar conditions for the English language. It is necessary a better alignment method to reduce the number of translations discarded. The testing results in HAREM, and translated data were similar and showed a clear difference to the current benchmark in both Portuguese and English. We believe that if a better token alignment solution is introduced in future work, the results obtained by this pipeline will be improved. However, as demonstrated in the following section dedicated to analysing the results obtained by the Portuguese NER models introduced in the library spaCy (5.6), even with all the limitations, our solution based on Ontonotes 5.0-PT performs better than this widely used NLP library.

5.5 Deployments to HuggingFace

The goal of PT-Pump-Up is to present solutions with higher benchmarking but, at the same time to increase the accessibility of off-the-shelf solutions. The outcome of our research resulted in matching the SOTA F1-Scores on HAREM selective and Default and introduced a new benchmark for ConLL-2003. In order to share these results with the community, we took advantage of the flexibility of our pipeline and integrated code to allow the deployment of our solutions to the Huggingface platform. As a result of our experiments, three new NER models are publically available in HuggingFace, the NER-PT-BERT-CRF-Conll2003[2], NER-PT-BERT-CRF-HAREM-Selective[3] and NER-PT-BERT-CRF-HAREM-Default[4].

The integration with Huggingface turned possible a solution that requires uniquely eight lines of Python code to execute. Furthermore, these models can be used as any other HuggingFace model, embedded in any preexistent code based on HF to pre-train, adaptation or deploy. We

[2] https://huggingface.co/arubenruben/NER-PT-BERT-CRF-Conll2003
[3] https://huggingface.co/arubenruben/NER-PT-BERT-CRF-HAREM-Selective
[4] https://huggingface.co/arubenruben/NER-PT-BERT-CRF-HAREM-Default

provide a quickstart Jupyter Notebook in the PT-Pump-Up official repository[5] where the code listed below is applied in practice.

Listing 5.1: Instructions to use our pipeline in a compact way.

```
from transformers import pipeline
import torch
import nltk

ner_classifier = pipeline(
    "ner",
    model="arubenruben/{REPLACE BY ONE OF THE PIPELINES}",
    device=torch.device("cuda:0") if torch.cuda.is_available() else torch.
    device("cpu"),
    trust_remote_code=True
)

text = "{INSERT TEXT TO BE CLASSIFIED HERE}"
tokens = nltk.wordpunct_tokenize(text)
result = ner_classifier(tokens)
```

In addition to models, all the datasets we used during our experiments were parsed, normalised to the BIO format and uploaded to HuggingFace. We share public access to Ontonotes 5.0-PT, First and Second HAREM in both selective, default and CoNLL-2003 style in a total of 10 new NER datasets ready for off-the-shelf usage inside the HuggingFace ecosystem without additional overhead.

5.6 Establish Performance Baselines using spaCy

During our research for Portuguese NER, we tested several training conditions that had not been previously benchmarked. We had, therefore, a comparability issue that limited the confidence in the results produced. In order to establish baselines that helped us measure the evolution of the work produced, we decided to develop an extra pipeline that operates in the spaCy ecosystem to extract the results of the Portuguese NER models included in this library in the scenarios we lacked comparability. spaCy is one of the most downloaded Python libraries and offers off-the-shelf solutions to several languages, including Portuguese. This library focuses on ease of use; spaCy performs poorly in NER challenges, yet, due to its simplicity, it is well known as a problem solver architecture. The authors of spaCy state they achieve a 0.90 F1-Score for Portuguese NER[6]. Our research proves that these results only stand in the Wikineural dataset.

5.6.1 Pipeline Architecture

The pipeline proposed has four steps. First, it loads the data from HuggingFace, converts it to the BIO notation, and then runs three parallel benchmarks in the test set. The results produced for

[5] https://github.com/arubenruben/PT-Pump-Up/blob/master/BERT-CRF.ipynb
[6] https://spacy.io/models/pt

pt_core_news_sm, *pt_core_news_md* and *pt_core_news_lg* are then plotted reusing code based on PyPlotly developed for the BERT-CRF pipeline.

5.6.2 spaCy NER Labelling Scheme

The NER model provided by spaCy is trained in a particular labelling scheme similar to Ontonotes 5.0. It includes the 10 BIO types we found in Ontonotes 5.0 (5.10) and extends it with an extra entity MISC.

An extra step of labelling normalisation is required to obtain benchmarks for the HAREM and CoNLL-2003 labelling schemes. The definition of rules to perform this conversion are specified in a configuration file without requiring modifying the source code. The conversions adopted are listed in Tables 5.13 and 5.14.

Table 5.13: HAREM to spaCy conversion map.

HAREM Label	**spaCy Label**
O	O
PER	PER
LOC	LOC
ORG	ORG
TEMPO	TIME
VALOR	QUANTITY
ABSTRACAO	MISC
ACONTECIMENTO	EVENT
COISA	PRODUCT
OBRA	WORK_OF_ART
OUTRO	MISC

Table 5.14: CoNLL-2003 to spaCy Conversion Map.

HAREM Label	**spaCy Label**
O	O
PER	PER
LOC	LOC
ORG	ORG
MISC	MISC

In the case of HAREM, the conversion to spaCy is lossy(e.g., VALOR can be either Money or Quantity). The CoNLL-2003 case is more straightforward, making direct mapping between label maps possible.

5.6.3 Zero-Shot Results Obtained

In this subsection, we summarize the F1-Scores obtained by spaCy in the same testing conditions used in the dataset combination (5.3) and data augmentation (5.4.4) sections of this document. These results are provided in Table 5.15.

Table 5.15: F1-Score comparative performance analysis between our BERT-CRF results(BERT-CRF) and spaCy Portuguese Large(PT-lg), Medium(PT-md) and Small(PT-sm) NER models in Portuguese NER datasets.

Test Set	**PT-sm**	**PT-md**	**PT-lg**	**BERT-CRF**
HAREM Default	0.368	0.396	0.414	0.787
HAREM Selective	0.394	0.421	0.442	0.832
HAREM CoNLL-2003	0.400	0.428	0.448	0.781
Wikineural	0.898	0.914	0.918	0.951
Ontonotes 5.0-PT	0.044	0.051	0.052	0.587
Ontonotes 5.0-PT (HAREM Selective)	0.242	0.282	0.285	0.558

The gains in F1-Score obtained in the medium and large NER models are marginal compared to the small case. None of the spaCy models can outperform our best BERT-CRF solutions in any labelling schemes. Excluding Wikineural, spaCy presents downgrades of 50% in the F1-Score when compared to our solutions.

5.7 Research Question Revisited

RQ1: Does combining datasets with the same labelling scheme but different sources positively impact the overall performance of models?

We demonstrate with the combination of First HAREM with Second HAREM that even datasets with the LS and the same annotators can quickly introduce entropy in the model. Making dataset combinations without further adjustments is incompatible with improvements on models' benchmarks.

RQ2: What is the additional overhead to introduce silver labelled data in the training pipeline?

Introducing data augmentation based on MT revealed a surprisingly challenging task due to the token alignment step. The results were poor and further work is necessary to ensure a viable tool of data augmentation based on MT for this NLP task.

RQ3: Does European Portuguese data negatively impact Pretrained Brazilian Portuguese transformer models?

It was impossible to obtain access to any of the LID models catalogued during our literature review (3.1). Without access to a preexistent LID tool and insufficient time to develop our solution, we were forced to abandon this experience for NER. This limitation is considered in the future work section of this document (7.4).

RQ4: Which are the best public framework to deploy an off-the-shelf NLP tool?

We decided to deploy our solutions uniquely in HuggingFace, due to our lack of time and because our quick study revealed that spaCy and stanza are still adopting their libraries to transformers architectures.

5.8 Summary

We selected NER as an essential NLP task to assess our methodology because it is a relevant upstream task that sustains many other High-Level NLP problems. Our literature review identified BERT-CRF as a promising architecture to use as a means to achieve new SOTA benchmarks in the three major LS used in Portuguese NER– Harem Default, Harem Selective and CoNLL-2003.

We began by combining and normalising several data sources into a single unit; using this approach, we obtained SOTA benchmarks for Portuguese NER in the dataset Wikineural, with 0.951 F1-Score. Nevertheless, due to the specificity of the annotation of HAREM, we could not further improve the current SOTA for Harem Default and Harem Selective.

Motivated by the need for improvement on HAREM, we introduced a pipeline of DA based on translation. We tested this pipeline by introducing a new Portuguese NER dataset, Ontonotes 5.0-PT, a translation of the English dataset Ontonotes 5.0. Token alignment was more challenging than expected; alignment tools output many incomplete results, negatively impacting the performance of the BERT-CRF training process using silver-labelled data.

Nevertheless, our spaCy baseline extraction pipeline demonstrates that all the solutions we developed and deployed on HuggingFace performed better than any of the three spaCy Portuguese NER models.

Chapter 6

Portuguese Abstractive Text Summarization

This chapter introduces the results obtained while researching Portuguese Abstractive Text Summarization. Here, we follow an improved methodology from the original proposal (4.2), based on the successes and failures of the previous chapter on Portuguese NER (5). The first section of this chapter clarifies the two experiments we developed to extend Portuguese resources for ATS (6.1). Then, we briefly introduce the Portuguese T5, the transformer architecture we established as a baseline to perform ATS due to the promising benchmarks shown (6.2). Then in Section 6.3, we cover the results obtained in the first experiment, evaluating similarity across translations from different platforms. Section 6.4 presents the ROUGE-L scores obtained by our new ATS training pipeline based on HuggingFace.

6.1 Redefining the Methodology

We used the conclusions established for NER to redefine our hypothesis before addressing Portuguese ATS. We identified two particular conclusions of the NER chapter relevant to transpose to ATS. The first relevant conclusion is that combining datasets tends to introduce levels of entropy that affect the model's training performances. The second is the need to better clarify if the bottleneck for data augmentation based on MT focuses uniquely on the alignment step.

These two conclusions motivated us to modify our approach to achieve SOTA results for Portuguese ATS. In this chapter, we do not perform any dataset combination step; in fact, we will explore even further DA based on MT, taking advantage of the fact that ATS does not require any post-translation alignment to demonstrate the limitations associated with augmented data faced previously are motivated exclusively by this step.

In order to study this subject deeper, we implemented not only an ATS training pipeline but also a dataset translation pipeline. We used this pipeline to translate the CNN-Dailymail dataset, the dataset we focus our attention on in the following subsections.

6.2 The T5-Model

This section describes the transformer-based architecture used during our research in Portuguese Abstractive TS. We selected T5 motivated by the promising results obtained by P. H. Paiola et al. [81] in this NLP task, following a methodology similar to the one proposed in this dissertation based on datasets combination, achieving a SOTA ROUGE-1 score of 49.91. Following a similar approach to the one used during the introduction of the BERT-CRF architecture, we focus not only on describing the T5 model (6.1) but also on presenting benchmarks obtained in other languages for ATS using this architecture (6.2.2)

6.2.1 Architecture

The T5 [92] is an encoder-decoder transformer model. Contrary to the BERT architecture that only includes the encoder side, the T5 model is more suitable for Text-2-Text NLP tasks due to the generative capabilities introduced by the decoder of this model.

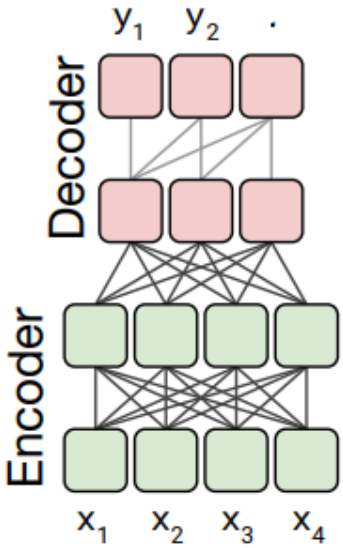


Figure 6.1: T5 encoder-decoder architecture [92].

In the context of ATS, the encoding-decoding problem is a mapping from **X** to **Y**, where **X** is an **N-Size** vector that represents the text to summarize and **Y** is an **M-Size** vector, that represents the result summary. In text summarization, the inequality $\mathbf{N} > \mathbf{M}$ stands.

The T5 architecture was initially proposed by Google in 2020 for the English language in three variations, small (60M Parameters), base(220M Parameters), and large (740M parameters) [92]. However, a Portuguese variant of this architecture was proposed due to the work of D. Carmo et al. [15]. The PT-T5 was trained in "the BrWac corpus, a large collection of web pages in Portuguese" [15], the same corpus as the Portuguese BERTimbau.

6.2.2 Results Obtained

In Table 6.1, we present the ROUGE-L SOTA benchmarks obtained by the T5 model for languages other than Portuguese. We aim to compare the actual Portuguese ATS benchmark with other languages to validate the results in Portuguese.

Table 6.1: Comparing the ROUGE-L performance of the T5 model in different languages.

Project Name	**Author**	**Language**	**T5**	**ROUGE-L**
ViT5: Pretrained Text-to-Text Transformer for Vietnamese Language Generation	Long Phan [85]	Vietnamese	Large	43.55
Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer [92]	Colin Raffel	English	Large	39.75
Deep Learning-Based Abstractive Summarization for Brazilian Portuguese Texts [81]	**Pedro H. Paiola**	**Brazilian Portuguese**	Base	**29.94**

The analysis of the results demonstrates that the PT-T5 scores lower ROUGE-L results than other languages using the same transformer architectures. This negative variation supports the need for further improvement in Portuguese ATS.

6.2.3 Training Limitations

As previously introduced in the subsection dedicated to the T5 architecture (6.2), this transformer is available in three variations, the small, base and large case. The difference between them is the number of parameters and the sequence input size accepted.

Abstractive Text Summarization is a more challenging task than NER; it requires more training and a larger volume of data to achieve an equivalent level of performance. Due to the scarcity of time and dedicated hardware resources to cover all the topics of this dissertation, we were forced to focus our attention on the small variation since it is the T5 case that requires less dedicated hardware and has faster training time due to the smaller number of gradients to back-propagate in each iteration.

6.3 Evaluating Similarity of Machine Translation Services

This section evaluates the similarity of the automatic machine translation services provided by three major IT companies, Google, Microsoft and Amazon Web Services. Our goal in this section is to collect more metrics that support our conclusion that the poor F1-Scores obtained while training NER models with silver-labelled data were associated not with the quality of the translation but rather with the alignment step.

First, we introduce the English dataset of ATS that we will use, the CNN-Dailymail (6.3.1). Then, we introduce the testing protocol (6.3.2). Subsection 6.3.3 presents the results obtained from

the experiments made. We conclude this section by exploring the challenges and the limitations faced while using these commercial MT solutions (6.3.4).

6.3.1 The CNN Dailymail Dataset

To validate our DA pipeline, we translated the CNN Dailymail [43, 106] dataset since it is the most used corpus in this field and is a benchmark for most of the literature. This dataset includes more than 300,000 news from CNN and Dailymail with 287,113 documents for training, 13,368 for validation and 11,490 for testing.

Due to the time and budget limitations faced (6.3.4), we decided to perform our experiments in a subset of this dataset composed of the first 10,000 training documents, the first 5,000 news of the validation set and the 10,000 documents of the testing set.

6.3.2 Experiments

We developed an additional pipeline to measure the document similarity across three Machine Translation systems– Google Translator, Microsoft Azure, Amazon Web Services(AWS) Translation Services and the variants of European and Brazilian Portuguese provided by some of these services. To assess this similarity, we measure the ROUGE-L between all pairs of document-document and summary-summary provided by these services. To explore all combinations, it would be necessary to perform twenty different experiments. Unfortunately due to the high cost of these translation services (6.3.4), it was only possible to execute two of the planned initially twenty experiments.

6.3.3 Results

Table 6.2 lists the ROUGE-L scores obtained between translation outputted by Google Translator and Azure PT-PT for both the documents and the summaries of CNN-Dailymail.

Table 6.2: Comparing the ROUGE-L of different commercial Machine Translation systems

Tool 1	**Tool 2**	**Training ROUGE-L**	**Validation ROUGE-L**	**Test ROUGE-L**	**Average ROUGE-L**
Google Translator Documents	Azure PT-PT Documents	13.12	12.63	12.93	12.93
Google Translator Summaries	Azure PT-PT Summaries	10.81	10.81	10.81	10.81

The results demonstrate a significant variation between the two translation engines. Unfortunately, we cannot conclude much from these results alone. We believe that if the original 20 experiments were performed, we could obtain a clear picture of the level of similarity across engines and their Portuguese variants.

6.3.4 Limitations

Commercial MT engines are typically integrated into Cloud Computing systems that are not specialised in MT. We found several outdated pieces of documentation pointing to unexistent topics. Furthermore, it is an expensive service; we estimate that translating the entire CNN-Dailymail dataset would cost around 18,000€. Some services offer free tiers but have extreme limitations on the daily rate of translations permitted. Lastly, these services are based on asynchronous translation jobs and do not guarantee to preserve the document's order. If order preservation is a requirement, additional solutions are required to overcome this limitation.

6.4 Abstractive Text Summarization Pipeline

This section summarises the work developed to create an ATS training pipeline based on the HuggingFace ecosystem. We took advantage of the work developed for NER (5.2.3) to reuse many software code, speeding the process of developing an ATS training solution. Our literature review did not identify an open-source ATS pipeline with sound software engineering standards that could be reused in this project.

Subsection 6.4.1 describes the features that our pipeline implements. Then, we present the results we obtained for Portuguese Abstractive TS (6.4.2) and list the limitations that these results present (6.4.3). Finally, we describe the contributions we uploaded to the HuggingFace hub (6.4.4).

6.4.1 Improving Code Accessibility

Motivated by the absence of an Abstractive Text Summarization Pipeline for the Portuguese language, we focus our attention on providing a flexible solution based on HuggingFace that would cover not only the essential features of training but also the capability of plotting the training stats using PyPlotly and offer automatic deployments to HuggingFace. The features implemented are summarised in Table 6.3. We are confident that the solution proposed simplifies the development of Abstractive TS since it is no longer necessary to produce a single line of code to train, test and deploy any HuggingFace-based ATS model; all relevant parameters are configurable using a configuration file.

Table 6.3: Summary of features implemented by our Abstractive Text Summarization pipeline.

Feature
Load From HuggingFace
Parameters Depend on Configuration File
Capacity to Pause Training
Integration with PyPlotly: Automatic Generation of Plots
Automatic Deployments to HuggingFace

6.4.2 Results

We performed three training processes in similar conditions using the PT-T5 architecture. First, we trained using the Portuguese subset of the XL-Sum dataset [40], then we focus on the training process using the two variations of the CNN-Dailymail-PT. Table 6.4 presents the results.

Table 6.4: Results of Abstractive Text Summarization models. Describing the training, validation and testing parameters. With focus on training dataset(Train), validation dataset(Val), test dataset(Test), the T5 model variant used(T5), learning rate (L.R), batch size(B.S), input length of documents(Len. D.) and summaries(Len. S.) and the ROUGE-L obtained in the test set.

Train	Val	Test	T5	L.R.	B.S	Len.D.	Len.S.	ROUGE-L
Portuguese XL-Sum	Portuguese XL-Sum	Portuguese XL-Sum	Small	1e-5	32	512	128	26.73
PT-CNN-Dailymail-Azure (10k)	PT-CNN-Dailymail-Azure (5k)	PT-CNN-Dailymail-Azure (10k)	Small	1e-5	32	512	128	25.23
PT-CNN-Dailymail-Google (10k)	PT-CNN-Dailymail-Google (5k)	PT-CNN-Dailymail-Google (10k)	Small	1e-5	32	512	128	25.56

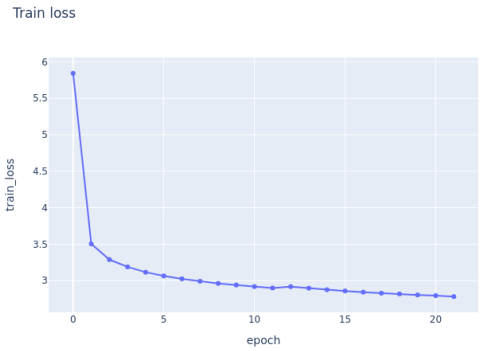


Figure 6.2: Training loss curve PT-CNN-Dailymail-Azure

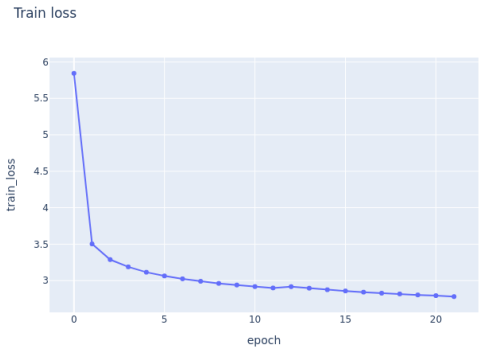


Figure 6.3: Training loss curve PT-CNN-Dailymail-Google

The models perform lower than expected. With these training parameters, we could not surpass the benchmark for Portuguese ATS. The training loss charts presented (6.2, 6.3) show an evident phenomenon of high bias of the T5 model used because the hyperbolic Cross-Entropy function stagnates in an undesirable high value of training loss. Despite the low ROUGE-L obtained, the proximity in the results offers two promising conclusions. Firstly, we believe introducing a larger T5, with more training parameters, would surpass the bias limitation demonstrated by the PT-T5-Small case. Secondly, the similar results between the Portuguese XL-Sum and our translation of the CNN-Dailymail dataset strengthen our confidence that it is possible to improve the results of several NLP tasks using augmented data based on MT services.

6.4.3 Limitations

Hardware limitations constrained the results presented in the subsections above. The inherent complexity of Abstractive Text Summarization requires extensive amounts of data to achieve exciting results. The bigger the dataset, the bigger the training time. This phenomenon is typically tackled with more powerful hardware, namely with the introduction of GPU parallelization; unfortunately, these resources are inaccessible due to budget constraints in this dissertation. This limited the integration of bigger variants of the T5-model; the inclusion of its base and large variations was unfeasible. Our intuition is that introducing these bigger architectures would improve the ROUGE-L results.

6.4.4 Deployments on HuggingFace

In this subsection, we present the public contributions in ATS uploaded to HuggingFace. We uploaded three summarization models to this platform, ptt5-portuguese-xlsum[1], ptt5-portuguese-cnn-daily-mail-azure[2] and ptt5-portuguese-cnn-daily-mail-google[3]. These resources were developed to be simple to use in few lines of code. Below we present the code guidelines on how to use this model (6.1).

Listing 6.1: Instructions to use the ATS model.

```
from transformers import T5Tokenizer, AutoModelForSeq2SeqLM

tokenizer = T5Tokenizer.from_pretrained('unicamp-dl/ptt5-base-portuguese-
    vocab')

model = AutoModelForSeq2SeqLM.from_pretrained("arubenruben/ptt5-portuguese-
    cnn-dailymail-azure-pt-pt")

text = "INSERT TEXT"

encoded_input = tokenizer(f"summarize: {text}", max_length=512, truncation=
    True, padding=True, return_tensors="pt")

summary = model(encoded_input)
```

6.5 Research Questions Revisited

RQ1: Does combining datasets from different sources positively impact the overall performance of models?

[1] https://huggingface.co/arubenruben/ptt5-portuguese-xlsum
[2] https://huggingface.co/arubenruben/ptt5-portuguese-cnn-daily-mail-azure
[3] https://huggingface.co/arubenruben/ptt5-portuguese-cnn-daily-mail-google

In this NLP task, we do not focus on answering this RQ. The results obtained in the context of NER are sufficient to clarify that this technique has several limitations. This chapter mainly focuses on consolidating our knowledge of DA techniques.

RQ2: What is the additional overhead to introduce augmented data in the training pipeline?

The similar values of ROUGE-L obtained in the same training conditions between original and silver-labelled data are promising. This result contrasts with the downgrade in F1-Score verified while training NER models with synthetic data. In the case of ATS, a token alignment step is not required, bypassing the limitations that negatively impact the results obtained for NER.

RQ3: Does European Portuguese data negatively impact Pretrained Brazilian Portuguese transformer models?

The results obtained for ATS in the T5-Small model do not permit extracting a conclusion on this topic. However, the similar results obtained by CNN-Dailymail translated in both variants of Portuguese do not present any sign that the European Portuguese variant is heavily impacted in its training process by a transformer architecture pre-trained in a Brazilian Portuguese corpus.

RQ4: Which are the best public framework to deploy an off-the-shelf NLP tool?

In the case of ATS, we do not modify the transformer architecture beyond the one already existing in the Huggingface hub. By avoiding external dependencies, any finetuned model based on Huggingface can be simply deployed in this platform. This motivated us to host the ATS results in HF, ensuring off-the-shelf solutions.

6.6 Summary

We decided to research Portuguese Abstractive Text Summarization due to the promising results of recent transformer architectures in this NLP task. Also, we needed an NLP task outside the token classification field to consolidate our knowledge in Machine Translation as a means to create silver-labelled data.

In this chapter, we created two pipelines, a pipeline of data augmentation based on Machine Translation and a training pipeline using the T5 transformer architecture. The DA pipeline is paramount to understand the research conducted in this chapter. We apply our best efforts to translate a subset of the English dataset CNN-Dailymail using several translation engines and explore the available solutions for European and Brazilian Portuguese. We intended to use this information to study the similarity in the output produced by these solutions and to clarify the impact of training Brazilian Portuguese pre-trained transformers with European Portuguese data. Unfortunately, due to the high cost of these services, our attempts were thwarted, and it was impossible to execute most of the initial experiments planned.

Secondly, we developed an ATS training pipeline to extract the results for three experiments. In the first experiment, we focused on validating the pipeline; we trained the Portuguese T5-Small transformer in the Portuguese XL-Sum dataset and compared it with the existent ROUGE-L benchmark. The results were similar. After ensuring the correctness of our solution, we finetune the model using our translation of the CNN-Dailymail dataset.

The results obtained were lower than the current benchmark, yet they are negatively influenced by bias in the model used. We believe that by introducing a T5 variation with more training parameters, we would obtain better results. Lastly, the ROUGE-L values obtained for the translated dataset are promising because they are similar to the results obtained in the gold-labelled dataset XL-Sum. This creates good perspectives that it is possible to achieve interesting results in several NLP tasks for low-resource languages like Portuguese using augmented data based on Machine Translation.

Chapter 7

Conclusions

This final chapter concludes the work developed in this dissertation. In Section 7.1, we reassert the most significant conclusions extracted from the work developed. Then, we focus on summarising the challenges we faced during our research (7.2) and listing the contributions introduced as a result of our work (7.3). We conclude by presenting topics of future work (7.4) that can be executed to extend the work started in this dissertation.

7.1 Conclusions

The recent advances in NLP based on Transformer architectures elevated the current SOTA benchmarks for several NLP tasks. This process was possible at the cost of large volumes of data and modern hardware resources. These requirements are acceptable for high-resource languages like English or Mandarin. However, in Portuguese, this is not the case. Working in low-resource languages like Portuguese introduces additional challenges due to the reduced amount of corpora and models available. These limitations tend to negatively impact the results that can be achieved in Portuguese NLP.

To address this problem, we follow a methodology to support the development of new Portuguese corpora and transformer-based models. This set of techniques not only enforces the development of new resources but also ensures that the resources produced can be used off-the-shelf, making them accessible to NLP researchers with any technical background. Regarding the scarcity of corpora, we propose a two-step methodology based on dataset combination, complemented by data augmentation founded on automatic machine translation. In addition, to address the need for more Portuguese models, we introduced training pipelines compatible with NER and ATS that take advantage of pre-trained transformer architectures to achieve satisfactory results in these Portuguese NLP tasks.

The process of validating this methodology required two steps. First, we were required to conduct an extensive literature review to identify which Portuguese NLP resources were available. We constrained our focus to eight NLP tasks, identifying fifty-nine datasets and thirty-three Portuguese models. In the second step, we decided which of the eight NLP tasks we should focus

during this dissertation. We selected Named Entity Recognition and Abstractive Text Summarization since they are two NLP tasks different in nature that demonstrated scarcity in the number of off-the-shelf resources available.

The research of NER and ATS answered all the previously formulated research questions (4.3). Firstly, the incapability of NER models to outperform the current SOTA using a process of dataset combination answers RQ1. Moreover, the challenges introduced by token alignments clarify RQ2. The complexity of introducing silver-labelled data is highly correlated with the nature of the NLP task. Our research proves that tasks not requiring a token alignment step offer lower overheads to include silver-labelled data.

Unfortunately, the limitations of accessing the LID models previously catalogued during our literature review limited our original research plan. It was impossible to split the corpora between European and Brazilian data sets. This fact thwarted the possibility of providing a clear answer to RQ3. However, the similar results obtained by our ATS models in European and Brazilian translations tend to conclude that the impact of mixing both variants of Portuguese on model training is reduced.

Finally, all the models and datasets developed are available in Huggingface. By taking advantage of the functionality introduced by this framework, it was possible to create off-the-shelf solutions that execute with four lines of Python code. The simplicity introduced by Huggingface provides a direct answer to RQ4.

7.2 Challenges

During our research in Portuguese NER and ATS, we faced several challenges that negatively influenced our work and opened new lines of future work worth mentioning. We have already listed those limitations separately within the chapters dedicated to those particular NLP tasks (5.4.5, 6.4.3). In this section, we recap them, highlighting the ones that impacted our work the most.

The NER research faced additional challenges motivated by the limitations of access to the LID models previously catalogued. This fact turned the inclusion of LID features within the training pipelines impracticable. Initially, we intended to actively separate European and Brazilian corpora to determine how different Portuguese variants impact the training process. Furthermore, the efforts to extend the existing NER corpora were negatively impacted by the low performance of token alignment tools. Both limitations created topics of future research that we intend to complete to improve the overall results obtained by PT-Pump-Up.

The development of ATS solutions also presented some unexpected challenges. We highlight the high cost of translation services provided by commercial tools as a significant limitation to our work. Initially, we intended to obtain a clear picture of how the existing machine translation engines perform, focusing on the differentiation these tools made across European and Brazilian Portuguese. However, these costs limited the number of experiments executed, denying the possibility of concluding anything relevant.

7.3 Contributions

We have already introduced an analysis of the contributions made to both NLP tasks (5.5, 6.4.4). Here, we recap those contributions and reinforce others that span across the entire project.

The research in Portuguese NER introduced three new off-the-shelf models based on the BERT-CRF architecture, one model for each of the three labelling schemes covered. Firstly, we highlight a new benchmark obtained for the Portuguese language with a 0.951 F1-Score in the Wikineural dataset. Secondly, we emphasise the translation of the NER dataset Ontonotes 5.0 and the normalisation of the labelling scheme of Second HAREM. Additionally, ATS research introduces three new models based on the T5 transformer and a new corpus composed of translated samples of the English dataset CNN-Dailymail. We use two different translation engines to obtain a European and a Brazilian version of the documents translated. All these contributions were developed with additional concerns to provide off-the-shelf solutions, being ready to use in HuggingFace[1].

During our research, we focused on ensuring that our solutions presented high levels of software engineering principles. The process of training new models and extending the existing corpora, previously described, required us to develop software components to support them. Therefore, for each of the NLP tasks, we developed two pipelines, one to support the training of models and another to allow the augmentation of the existing corpora. All the code developed is open-source and publicly available in the PT-Pump-Up GitHub[2] repository.

Finally, the extensive resource cataloguing of NLP resources made during our literature review process is publically available in the PT-Pump-Up GitHub repository.

7.4 Future Work

We identify ten lines of research for further development of PT-Pump-Up, we list all the possibilities in the following sections, and we conclude by explaining which ones we consider the most prioritised (7.4.10).

7.4.1 Deploy Results in Portulan Clarin

As mentioned, the results obtained during this dissertation are deployed in Huggingface and GitHub. We selected these platforms because they provide features of automatic deployment that can be integrated with our pipelines. However, these platforms are not focused on the Portuguese language; during our literature review, we identified Portulan Clarin[3] as the primary hub for Portuguese NLP. Unfortunately, the current release of this platform does not offer automatic deployments, relying on manual submissions. As future work, we believe it will be worth publishing our results on this platform even though it is a more time-consuming task.

[1] https://huggingface.co/arubenruben
[2] https://github.com/arubenruben/PT-Pump-Up
[3] https://portulanclarin.net/

7.4.2 Extend the Methodology to other NLP Tasks

The two NLP tasks we focus on are a small subset of all the universe of Portuguese NLP tasks that require a high level of dedication in order to provide the community with results similar in performance with high resourced languages. The idea is, therefore, scale the same methodology to other NLP tasks in order to increase the public resources available for them.

7.4.3 Create a Language Identification Model

As previously mentioned, our original goal of including capabilities of differentiating European and Brazilian Portuguese in our training pipelines was thwarted by the incapacity to gain access to those models. This fact creates an interest in developing an off-the-shelf, easy-access solution for this NLP task within the context of the PT-Pump-Up project.

7.4.4 Redefine the Training Parameters for Existing Pipelines

This dissertation was not about showing extensive knowledge of how complex transformers models behave. Our intuition is that if we focus more time on better understanding them, we will find a better set of training parameters to unlock the results obtained even further.

7.4.5 Improve the Accessibility to the results of our Literature Revision

As mentioned, we consider the information we aggregated during our research on SOTA NLP resources in the Portuguese Language to be precious to simplify access to Portuguese NLP and motivate other personnel to contribute. The deployment is currently a simple markdown document with references to those resources. However, a web interface is possible and desirable. A cleaner and more modern layout for PT-Pump-Up would decrease the technological overhead of accessing this information, motivating others to participate.

7.4.6 Further Improvements on Abstractive TS Pipeline

The results for Abstractive TS were unsatisfactory, performing lower than expected. Unlike NER, where we could train models with good performance levels, the lack of time to wait extensive training periods and the focus on studying the behaviour of automatic machine translation tools motivated these lower results. More time and better hardware will be required to handle this task properly.

7.4.7 Establish a Better Metric to Evaluate the Similarity between MT Engines

The Results of the similarity experiment summarised in Table 6.3.3 revealed inconclusive. Due to the limitations imposed by Microsoft and AWS, the experiment was a failure. We strongly believe that if all the combinations of engines initially planned were made, the result would be different, and the extraction of clear insight would be possible. Nevertheless, the two experiments revealed

a low value of ROUGE-L between them, which indicates that maybe this metric is not the most suitable to evaluate the similarity between outputs. Future work could explore other metrics based on cosine similarity, BLUE or other more refined SOTA techniques to extract valid conclusions from this experiment.

7.4.8 Extend Deployments to spaCy and Stanza

Integrating our solutions in major NLP frameworks other than HuggingFace is interesting from the point of view of accessibility. Most people use spaCy and will not stop, even if our solution is comparably better for the Portuguese case. Therefore the only way to provide the community with better models is to take advantage of their open-source policies to request the submission of our solutions into the ecosystems of these platforms.

7.4.9 Refactor the Token Alignment Functionality. Introduce a Validation metric

The attempt to augment the HAREM dataset using machine translation-based DA techniques revealed unexpected challenges due to the lack of good-quality alignment tools for the Portuguese language. These constraints heavily limited the results obtained while translating the dataset Ontonotes 5.0 to Portuguese; compared with English, a similar training process executed in this dataset performs 0.40 less F1-Score. This shows how incomplete these alignments are and the need for further work to be developed.

Furthermore, all the evaluations of the alignments produced were performed in the downstream NLP task. In the case of NER, we evaluate the quality of Ontonotes 5.0 based on the F1-Score obtained by the model in a gold-labelled test set. We identify two reasons this evaluation technique is undesirable. Firstly, it does not objectively measure how good the alignments are since it is impacted by the bias associated with the downstream task. Secondly, it requires a prior training process, which is often time and computation-consuming. Therefore, a more straightforward solution is required to evaluate the alignments immediately.

Despite the limitations, this technique provides some insights into the quality of the alignments when performing a direct comparison with the English case in similar training conditions is possible.

7.4.10 Prioritising Future Work: Token Alignment and Stanza Deployments

Creating a good token alignment tool between Portuguese and English is paramount to unlocking many data augmentation possibilities for Token Classification tasks. It should, therefore, be prioritised. Also, deploying on Stanza should be a relevant effort to execute; this NLP framework, unlike spaCy, has no implementations for the Portuguese language yet. Introducing better Portuguese NER and POS tagging models in this ecosystem would be a valuable contribution to both projects.

References

[1] Natural language processing [nlp] market size: Growth 2029. https://www.fortunebusinessinsights.com/industry-reports/natural-language-processing-nlp-market-101933. Accessed: July 4, 2023.

[2] Portuguese - worldwide distribution. https://www.worlddata.info/languages/portuguese.php. Accessed: July 4, 2023.

[3] Susana Afonso, Eckhard Bick, Renato Haber, and Diana Santos. Floresta sintá (c) tica: a treebank for portuguese. In *quot; In Manuel González Rodrigues; Carmen Paz Suarez Araujo (ed) Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)(Las Palmas de Gran Canaria Espanha 29-31 de Maio de 2002) Paris: ELRA*. ELRA, 2002.

[4] Ēriks Ajausks, Victoria Arranz, Laurent Bié, Aleix Cerdà-i Cucó, Khalid Choukri, Montse Cuadros, Hans Degroote, Amando Estela, Thierry Etchegoyhen, Mercedes García-Martínez, et al. The multilingual anonymisation toolkit for public administrations (mapa) project. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 471–472, 2020.

[5] Miguel B. Almeida, Mariana S. C. Almeida, André F. T. Martins, Helena Figueira, Pedro Mendes, and Cláudia Pinto. Priberam compressive summarization corpus: A new multi-document summarization corpus for European Portuguese. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 146–152, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[6] Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998.

[7] Florbela Barreto, António Branco, Eduardo Ferreira, Amália Mendes, Maria Fernanda Bacelar do Nascimento, Filipe Nunes, and João Ricardo Silva. Open resources and tools for the shallow processing of portuguese: the tagshare project. In *Proceedings of the V International Conference on Language Resources and Evaluation-LREC2006*. European Language Resources Association, 2006.

[8] David Soares Batista, David Forte, Rui Silva, Bruno Martins, and Mário Silva. Extracção de relações semânticas de textos em português explorando a dbpédia e a wikipédia. *Linguamática*, 5(1):41–57, Jul. 2013.

[9] António Branco, Catarina Carvalheiro, Sílvia Pereira, Sara Silveira, João Ricardo Silva, Sérgio Castro, and João Graça. A propbank for portuguese: the cintil-propbank. In *LREC*, pages 1516–1521, 2012.

[10] António Branco and João Ricardo Silva. Evaluating solutions for the rapid development of state-of-the-art pos taggers for portuguese. In *LREC*, 2004.

[11] Henrique Lopes Cardoso. Language models, Apr 2022.

[12] Henrique Lopes Cardoso. Nlp intro, Feb 2022.

[13] Henrique Lopes Cardoso. Sequence labeling, Apr 2022.

[14] Paula CF Cardoso, Erick G Maziero, Mara Luca Castro Jorge, Eloize MR Seno, Ariani Di Felippo, Lucia Helena Machado Rino, Maria das Gracas Volpe Nunes, and Thiago AS Pardo. Cstnews-a discourse-annotated corpus for single and multi-document summarization of news texts in brazilian portuguese. In *Proceedings of the 3rd RST Brazilian Meeting*, pages 88–105, 2011.

[15] Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data, 2020.

[16] Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*, 2020.

[17] Dayvid W. Castro, Ellen Souza, Douglas Vitório, Diego Santos, and Adriano L.I. Oliveira. Smoothed n-gram based models for tweet language identification: A case study of the brazilian and european portuguese national varieties. *Applied Soft Computing*, 61:1160–1172, 2017.

[18] Papers With Code. Data-to-text generation.

[19] Papers With Code. Information retrieval.

[20] Papers With Code. Language identification.

[21] Sandra Collovini, Thiago I Carbonel, Juliana Thiesen Fuchs, Jorge César Coelho, Lúcia Rino, and Renata Vieira. Summ-it: Um corpus anotado com informaç oes discursivas visandoa sumarizaç ao automática. *Proceedings of TIL*, 121, 2007.

[22] Sandra Collovini, Joaquim Francisco Santos Neto, Bernardo Scapini Consoli, Juliano Terra, Renata Vieira, Paulo Quaresma, Marlo Souza, Daniela Barreiro Claro, and Rafael Glauber. Iberlef 2019 portuguese named entity recognition and relation extraction tasks. In *IberLEF@ SEPLN*, 2019.

[23] Roque Lopez Condori, Thiago Pardo, Lucas Avanço, Pedro Balage Filho, Alessandro Bokan, Paula Cardoso, Márcio Dias, Fernando Nóbrega, Marco Cabezudo, Jackson Souza, et al. A qualitative analysis of a corpus of opinion summaries based on aspects. In *Proceedings of the 9th Linguistic Annotation Workshop*, pages 62–71, 2015.

[24] Francisco Costa and António Branco. Lx-timeanalyzer: A temporal information processing system for portuguese. 2012.

[25] Francisco Costa and António Branco. Timebankpt: a timeml annotated corpus of portuguese. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3727–3734, 2012.

[26] Luís Filipe Cunha, Ricardo Campos, and Alípio Jorge. Event extraction for portuguese: A qa-driven approach using ace-2005. In *Springer's LNAI – Lecture Notes in Artificial Intelligence*. Springer, 2023.

[27] Joaquim Ferreira Da Silva and Gabriel Pereira Lopes. Identification of document language is not yet a completely solved problem. In *2006 International Conference on Computational Inteligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (CIMCA'06)*, pages 212–212, 2006.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[29] Zi-Yi Dou and Graham Neubig. Word alignment by fine-tuning embeddings on parallel corpora. In *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2021.

[30] Magali Sanches Duran and Sandra Maria Aluísio. Propbank-br: a brazilian treebank annotated with semantic role labels. In *LREC*, pages 1862–1867, 2012.

[31] Chris Dyer, Victor Chahuneau, and Noah A Smith. A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, 2013.

[32] Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165:113679, 2021.

[33] Claire Cardie Faisal Ladhak, Esin Durmus and Kathleen McKeown. Wikilingua: A new benchmark dataset for multilingual abstractive summarization. In *Findings of EMNLP, 2020*, 2020.

[34] Diego de Vargas Feijo and Viviane P Moreira. Improving abstractive summarization of legal rulings through textual entailment. *Artificial Intelligence and Law*, pages 1–23, 2021.

[35] Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard H. Hovy. A survey of data augmentation approaches for NLP. *CoRR*, abs/2105.03075, 2021.

[36] Evandro B Fonseca, André Antonitsch, Sandra Collovini, Daniela Amaral, Renata Vieira, and Anny Figueira. Summ-it++: an enriched version of the summ-it corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2047–2051, 2016.

[37] Cláudia Freitas, Paula Carvalho, Hugo Gonçalo Oliveira, Cristina Mota, and Diana Santos. Second harem: advancing the state of the art of named entity recognition in portuguese. In *quot; In Nicoletta Calzolari; Khalid Choukri; Bente Maegaard; Joseph Mariani; Jan*

Odijk; Stelios Piperidis; Mike Rosner; Daniel Tapias (ed) Proceedings of the International Conference on Language Resources and Evaluation (LREC 2010)(Valletta 17-23 May de 2010) European Language Resources Association. European Language Resources Association, 2010.

[38] Cláudia Freitas, Diana Santos, Cristina Mota, Hugo Gonçalo Oliveira, and Paula Carvalho. Relation detection between named entities: report of a shared task. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 129–137, 2009.

[39] Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732, 2022.

[40] Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online, August 2021. Association for Computational Linguistics.

[41] Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 473–483, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[42] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517, 2016.

[43] Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *NIPS*, pages 1693–1701, 2015.

[44] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. Knowledge graphs. *ACM Comput. Surv.*, 54(4), jul 2021.

[45] Shulin Hu, Huajun Zhang, Xuesong Hu, and Jinfu Du. Chinese named entity recognition based on bert-crf model. In *2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS)*, pages 105–108, 2022.

[46] IBM. What is natural language processing? https://www.ibm.com/topics/natural-language-processing, 2022. Accessed: July 4, 2023.

[47] Alankar Jain, Bhargavi Paranjape, and Zachary C Lipton. Entity projection via machine translation for cross-lingual ner. *arXiv preprint arXiv:1909.05356*, 2019.

[48] Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782, 2019.

[49] Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[50] Dan Jurafsky and James H. Martin. Naive bayes and sentiment classification. In *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, pages 1–1. Pearson, Noida, 2022.

[51] Dan Jurafsky and James H. Martin. Regular expressions, text normalization, edit distance. In *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, pages 13–13. Pearson, Noida, 2022.

[52] Dan Jurafsky and James H. Martin. Relation and event extraction. In *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*, page 1–12. Pearson, 2022.

[53] Dan Jurafsky and James H. Martin. Semantic role labeling. In *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson, Noida, 2022.

[54] Paul R Kingsbury and Martha Palmer. From treebank to propbank. In *LREC*, pages 1989–1993, 2002.

[55] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86, 2005.

[56] Joos Korstanje. The f1 score. https://towardsdatascience.com/the-f1-score-bec2bbc38aa6, Aug 2021. Accessed: July 4, 2023.

[57] Lewis. What is the rouge metric? https://www.youtube.com/watch?v=TMshhnrEXlg&ab_channel=HuggingFace, 2023. Accessed: July 4, 2023.

[58] Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*, 2019.

[59] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[60] Hui Lin and Vincent Ng. Abstractive summarization: A survey of the state of the art. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9815–9822, 2019.

[61] Pierre Lison and Jörg Tiedemann. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA).

[62] Alexandre Lopes, Rodrigo Nogueira, Roberto Lotufo, and Helio Pedrini. Lite training strategies for portuguese-english and english-portuguese translation. *arXiv preprint arXiv:2008.08769*, 2020.

[63] Fábio Lopes, César Teixeira, and Hugo Gonçalo Oliveira. Contributions to clinical named entity recognition in Portuguese. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 223–233, Florence, Italy, August 2019. Association for Computational Linguistics.

[64] Pedro H. Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Lecture Notes on Computer Science (LNCS), pages 313–323, Canela, RS, Brazil, September 24-26 2018. Springer.

[65] Bruno Martins and Mário J. Silva. Language identification in web pages. In *Proceedings of the 2005 ACM Symposium on Applied Computing*, SAC '05, page 764–768, New York, NY, USA, 2005. Association for Computing Machinery.

[66] Rada Mihalcea and Paul Tarau. A language independent algorithm for single and multiple document summarization. In *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*, 2005.

[67] Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*, 2021.

[68] Diego Mollá, Menno Van Zaanen, and Daniel Smith. Named entity recognition for question answering. In *Proceedings of the Australasian language technology workshop 2006*, pages 51–58, 2006.

[69] Cristina Mota and Diana Santos. Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O segundo harem, 2008.

[70] Diego Moussallem, Mohamed Ahmed Sherif, Diego Esteves, Marcos Zampieri, and Axel-Cyrille Ngonga Ngomo. LIDIOMS: A multilingual linked idioms data set. *CoRR*, abs/1802.08148, 2018.

[71] Andrew Ng. Bleu score. https://youtu.be/DejHQYAGb7Q, Jan 2023. Accessed: July 4, 2023.

[72] Jun-Ping Ng and Viktoria Abrecht. Better summarization evaluation with word embeddings for ROUGE. *CoRR*, abs/1508.06034, 2015.

[73] Fernando Antônio Asevedo Nóbrega, Thiago Alexandre Salgueiro Pardo, and Núcleo Interinstitucional de Linguística Computacional. Rearrangement and creation of new corpora for update and compressive summarization tasks for portuguese language. *Caderno de resumos*, 2017.

[74] Joel Nothman, Nicky Ringland, Will Radford, Tara Murphy, and James R. Curran. Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, 194:151–175, 2013. Artificial Intelligence, Wikipedia and Semi-Structured Resources.

[75] Ana Sofia Medeiros Oliveira. Semantic role labeling in portuguese: improving the state of the art with transfer learning and bert-based models. Master's thesis, FCUP, 2020.

[76] André Seidel Oliveira and Anna Helena Reali Costa. Plsum: Generating pt-br wikipedia by summarizing multiple websites. *arXiv preprint arXiv:2112.01591*, 2021.

[77] Hugo Gonçalo Oliveira, Paulo Gomes, Nuno Seco, and Diana Santos. *PAPEL*.

[78] Hugo Gonçalo Oliveira, Leticia Antón Pérez, Hernani Costa, and Paulo Gomes. Uma rede léxico-semântica de grandes dimensões para o português, extraída a partir de dicionários electrónicos. *Linguamática*, 3(2):23–38, 2011.

[79] Sofia Oliveira, Daniel Loureiro, and Alípio Jorge. Improving portuguese semantic role labeling with transformers and transfer learning. In *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–9. IEEE, 2021.

[80] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[81] Pedro H Paiola, Gustavo H de Rosa, and João P Papa. Deep learning-based abstractive summarization for brazilian portuguese texts. In *Intelligent Systems: 11th Brazilian Conference, BRACIS 2022, Campinas, Brazil, November 28–December 1, 2022, Proceedings, Part II*, pages 479–493. Springer, 2022.

[82] Martha Palmer, Daniel Gildea, and Paul Kingsbury. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, 03 2005.

[83] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[84] Thiago Alexandre Salgueiro Pardo and Lucia Helena Machado Rino. Temário: Um corpus para sumarização automática de textos. *São Carlos: Universidade de São Carlos, Relatório Técnico*, 2003.

[85] Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. ViT5: Pretrained text-to-text transformer for Vietnamese language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 136–142, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics.

[86] André Ricardo Oliveira Pires. Named entity extraction from portuguese web text. 2017.

[87] Elvys Linhares Pontes, Juan-Manuel Torres-Moreno, Stéphane Huet, and Andréa Linhares. A new annotated portuguese/spanish corpus for the multi-sentence compression task. In *11th International Conference on Language Resources and Evaluation (LREC)*, 2018.

[88] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Joint conference on EMNLP and CoNLL-shared task*, pages 1–40, 2012.

[89] James Pustejovsky. Timebank 1.2. *http://www. ldc. upenn. edu/*, 2006.

[90] James Pustejovsky, José Castano, Robert Ingria, Roser Saurı, Rob Gaizauskas, Andrea Setzer, Graham Katz, and D Radev. Timeml: A specification language for temporal and event expressions. In *Proceedings of the International Workshop of Computational Semantics*, page 193, 2003.

[91] Libo Qin, Minheng Ni, Yue Zhang, and Wanxiang Che. Cosda-ml: Multi-lingual code-switching data augmentation for zero-shot cross-lingual nlp. *arXiv preprint arXiv:2006.06402*, 2020.

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

[93] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*, 2018.

[94] EHUD REITER and ROBERT DALE. Building applied natural language generation systems. *Natural Language Engineering*, 3(1):57–87, 1997.

[95] Rafael Ribaldo, Ademar Takeo Akabane, and Thiago Alexandre Salgueiro Pardo. Multi-document summarization with graph metrics. 2012.

[96] Eric Roberts. Nlp overview. https://cs.stanford.edu/people/eroberts/courses/soco/projects/2004-05/nlp/overview_history.html, 2004.

[97] Roberts Rozis and Raivis Skadiņš. Tilde model-multilingual open data for eu languages. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 263–265, 2017.

[98] Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela. HAREM: An advanced NER Evaluation Contest for Portuguese. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, 22-28 May 2006. European Language Resources Association (ELRA).

[99] Joaquim Santos, Bernardo Consoli, Cicero dos Santos, Juliano Terra, Sandra Collonini, and Renata Vieira. Assessing the impact of contextual embeddings for portuguese named entity recognition. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 437–442, 2019.

[100] Joaquim Santos, Bernardo Consoli, Cicero dos Santos, Juliano Terra, Sandra Collonini, and Renata Vieira. Assessing the impact of contextual embeddings for portuguese named entity recognition. In *Proceedings of the 8th Brazilian Conference on Intelligent Systems*, pages 437–442, 2019.

[101] Rodrigo Santos, João Silva, António Branco, and Deyi Xiong. The direct path may not be the best: Portuguese-chinese neural machine translation. In *EPIA Conference on Artificial Intelligence*, pages 757–768. Springer, 2019.

[102] Luís Sarmento. O siemês e a sua participação no harem e no mini-harem. *quot; In Diana Santos; Nuno Cardoso (ed) Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM a primeira avaliação conjunta na área Linguateca 2007*, 2007.

[103] Phillip Schneider, Tim Schopf, Juraj Vladika, Mikhail Galkin, Elena Simperl, and Florian Matthes. A decade of knowledge graphs in natural language processing: A survey. 09 2022.

[104] Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia. *CoRR*, abs/1907.05791, 2019.

[105] developers scikit learn. 6.1. pipelines and composite estimators. https://scikit-learn.org/stable/modules/compose.html#pipeline.

[106] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[107] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany, August 2016. Association for Computational Linguistics.

[108] Sara Botelho Silveira and António Branco. Extracting multi-document summaries with a double clustering approach. In *International Conference on Application of Natural Language to Information Systems*, pages 70–81. Springer, 2012.

[109] SmartDataAnalytics. Knowledge-graph-analysis-programming-exercises/readme.md at master · smartdataanalytics/knowledge-graph-analysis-programming-exercises, Jan 2018.

[110] Afonso Sousa, Bernardo Leite, Gil Rocha, and Henrique Lopes Cardoso. Cross-lingual annotation projection for argument mining in portuguese. In *Progress in Artificial Intelligence: 20th EPIA Conference on Artificial Intelligence, EPIA 2021, Virtual Event, September 7–9, 2021, Proceedings*, page 752–765, Berlin, Heidelberg, 2021. Springer-Verlag.

[111] Hugo Sousa, Alípio Jorge, and Ricardo Campos. tieval: An evaluation framework for temporal information extraction systems. *arXiv preprint arXiv:2301.04643*, 2023.

[112] Hugo Sousa, Alipio Mario Jorge, Arian Pasquali, Catarina Santos, and Mario Lopes. A biomedical entity extraction pipeline for oncology health records in portuguese. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*, SAC '23, page 950–956, New York, NY, USA, 2023. Association for Computing Machinery.

[113] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*, 2019.

[114] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. Bertimbau: Pretrained bert models for brazilian portuguese. In Ricardo Cerri and Ronaldo C. Prati, editors, *Intelligent Systems*, pages 403–417, Cham, 2020. Springer International Publishing.

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

[116] Jannik Strötgen and Michael Gertz. A baseline temporal tagger for all languages. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 541–547, 2015.

[117] Ann Taylor, Mitchell Marcus, and Beatrice Santorini. The penn treebank: An overview. 01 2003.

[118] Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics:*

EMNLP 2021, pages 2521–2533, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[119] Jörg Tiedemann. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248, 2009.

[120] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003.

[121] Fabricio E da S Tosta, Ariani Di Felippo, and Thiago AS Pardo. Estudo de métodos clássicos de sumarização no cenário multidocumento multilíngue. In *STUDENT WORKSHOP ON INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (TILiC)*, volume 3, pages 1–3, 2013.

[122] Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, 2013.

[123] Gael Vaamonde. Ps post scriptum: Dos corpus diacrónicos de escritura cotidiana. *Procesamiento del lenguaje natural*, (55):57–64, 2015.

[124] Diego de Vargas Feijó and Viviane Pereira Moreira. Rulingbr: A summarization dataset for legal texts. In *International Conference on Computational Processing of the Portuguese Language*, pages 255–264. Springer, 2018.

[125] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[126] Renata Vieira, Fernanda Olival, Helena Cameron, Joaquim Santos, Ofélia Sequeira, and Ivo Santos. Enriching the 1758 portuguese parish memories (alentejo) with named entities. *Journal of Open Humanities Data*, 7:20, 2021.

[127] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6256–6268. Curran Associates, Inc., 2020.

[128] Marcos Zampieri and Binyam Gebrekidan Gebre. Automatic identification of language varieties: The case of portuguese. In *KONVENS2012-The 11th Conference on Natural Language Processing*, pages 233–237. Österreichischen Gesellschaft für Artificial Intelligende (ÖGAI), 2012.

[129] Xiang Zhang and Yann LeCun. Text understanding from scratch, 2015.

Appendix A

T5-CRF Pipeline for Named Entity Recognition

Motivated by the limitations in further improving the benchmarks for NER in HAREM Selective and Default, we decided to explore alternative architectures to enhance the training process. Due to the limited availability of Portuguese Transformers, we opted for the T5 architecture to validate our hypothesis.

The PT-T5 [15] follows the original T5 architecture and is categorized as an Encoder-Decoder architecture. To extract F1-Score measures using this transformer, we excluded the decoder component of the architecture and substituted it with our Conditional Random Field.

Compared to BERT, the PT-T5 transformer is more recent and supports larger input sizes. Table A.1 presents the results obtained during our training process.

Table A.1: Results of T5 pipeline. Describing all the training parameters, the Dataset (T.D), the T5 model used (T5), the Learning Rate (L.R), the batch size (B.S), the Input Sequence Length (Seq Len), and the F1-Score.

T.D	**T5**	**L.R**	**B.S**	**Seq Len**	**F1-Score**
P. HAREM	Base	1e-5	4	1024	0.800
P. HAREM	Large	1e-5	4	1024	0.773

The T5-Model did not surpass the performance of the BERT-CRF architecture, which is why we included this information in an appendix. Since there was no significant improvement, it was not the primary focus of our attention. However, a positive outcome of exploring this research direction was the increased flexibility it brought to our original pipeline. With minor modifications to the existing code, we were able to train models from the T5 transformer family. This demonstrates that our code is not overfitted to a specific architecture, unlike many other available solutions.

Appendix B

Transpiling Portuguese Natural Language Processing Chapter to Markdown

In order to quickly publish our survey on Portuguese NLP Resources in an accessible format in GitHub[1], we leveraged the excellent transpilation capabilities of ChatGPT[2] to convert LaTeX to Markdown. By utilizing this feature, we were able to provide a solution within the document's release timeline. We maintained all prompts within the same conversation thread to provide contextualization to the model, and we performed thorough visual validation of the obtained results. Fortunately, no issues arose during this process.

It's important to note that the assessment of ChatGPT's transpilation capabilities is not the main focus of our dissertation; hence, this information is included in the appendix.

[1] https://github.com/arubenruben/PT-Pump-Up

[2] https://chat.openai.com/share/d5713c6e-d1fa-46d6-95d3-f990e0cb6d76