Named entity recognition specialised for Portuguese 18^{th}-century History research

Joaquim Santos[1], **Renata Vieira**[2], **Fernanda Olival**[2], **Helena Freire Cameron**[3], **Fátima Farrica**[2]
[1]University of Vale do Rio dos Sinos, Brazil
[2]CIDEHUS - University of Évora, [3]CIDEHUS - Portalegre Polytechnic University, Portugal
nejoaquim@edu.unisinos.br, renatav@uevora.pt,
mfo@uevora.pt, helenac@ipportalegre.pt,fatimafarrica@sapo.pt

Abstract

This paper presents the construction of a corpus and the respective models learned for the Named Entity Recognition (NER) task, specialised for historical research. The entity categories were adapted based on the objectives of the historical analysis of the 18th-century text. We trained and evaluated traditional neural networks and the new Large Language Models (LLMs) for the NER task. In total, we assessed six language models, where the results of traditional architectures were superior to LLMs.

1 Introduction

This work presents a study performed on a collection of historical Portuguese texts called the *Parish Memories* produced between 1758-1761. The texts have been manually transcribed and normalised. The study involves i) the definition of a special set of entity categories for annotation based on the expertise of historians, ii) manual annotation of a subset of this collection, and iii) the evaluation of machine learning models for the task of annotation of these categories.

Previously trained systems for Named Entity Recognition (NER) cannot be applied here, as we used a distinct set of categories, and they differ in various ways from the usual ones. Therefore, we needed to adapt the training to build new Machine Learning (ML) models. We made use of previously studied configurations (Santos et al., 2019) to train the models, and also considered alternative options with more recently available language models.

The goal is to apply the best models in the future to help in the annotation process of the whole historical collection. With the results we achieved, we believe it will be possible to use the models through an assisted-based semi-automated annotation system.

2 Related Work

The task of Named Entity Recognition is a highly studied task, and there are many works devoted to the Portuguese language. However, it is more common to find works related to contemporary Portuguese. A recent survey on NER for contemporary Portuguese is presented in (Albuquerque et al., 2023).

NER for historical Portuguese texts are more difficult to find. There are similar studies made for other languages, in (Ehrmann et al., 2023) we find a survey on Named Entity Recognition and Classification in Historical Documents. This survey refers to the Portuguese Historical Corpus, BDCamões (Grilo et al., 2020). This *corpus* was automatically annotated with natural language processing tools, includes the usual categories of NE, and there is no evaluation of the accuracy of the annotation performed.

In our case, we are studying a Portuguese historical corpus from the 18^{th} century annotated with historical-oriented subcategories. We present an evaluation of the accuracy of current models based on the dataset that was manually annotated.

By the nature of this particular corpus, by its linguistic and historical value, and the plurality of authors that wrote the *Parish Memories*, we consider that it can be helpful not only for historians and linguists, but also for architects, demographers, territory administrators, and planners.

3 Historical source: the *Parish Memories* Corpus

The *Parish Memories* are the answers to a survey with 60 questions sent in January of 1758 to the bishops asking them to resend it to the parish priests of the entire kingdom of Portugal to respond to it. The inquiry has two main goals: 1) to obtain feedback about the state of the territory after the big earthquake of 1755; 2) to gather information to create a Geographical Dictionary of Portugal.

Nowadays, on the Portuguese National Archive of Torre do Tombo website, the Parish Memories' manuscripts are available online as digitised copies from microfilms. In this work, we consider a subset from the biggest region of Portugal (Alentejo). The originals have been manually transcribed, normalised and annotated with named entities.

In previous work (Vieira et al., 2021), we have performed experiments with three basic categories (PERSON, LOCAL, ORGANISATION) and then we performed a *corpus*-based study to define the extension of these categories (Cameron et al., 2022).

4 Manual annotation of the historical source

4.1 NE categories customized to History research

Our recent annotation process tries to translate the complexity of past ages expressed in historical sources, as they differ from contemporary ones.

We started by considering five main categories: PERSON, PLACE, ORGANISATION, TIME and AUTHOR WORK. The first four aim to respond to historical questions: Who, Where, What, When, and the last allows us to treat the text sources mentioned in the *corpus*.

The main categories PERSON and PLACE were broken down into several subcategories due to their complexity and according to their relevance to the study of the source.

The category person (PER) considers references by name, occupation, or social category (in that order of preference if more than one appears in the expression). Also, we defined specific subcategories for mentions of saints, divinities, groups of persons, and authors. Examples of mentions to persons by occupation are:

- Arcebispo de Évora [Archbishop of Évora]
- Presidente da Mesa da Consciência [President of the Military Orders Council]

An example of a social category is Conde da Torre [Count of the Tower]. The subcategory for groups of persons is used to annotate organic groups, families and members of an organisation, among others, as seen in the following examples:

- Jesuítas [the Jesuits]
- Sequeiras [the Sequeira family]
- Almas [Souls]
- Mouros [the Moors]

Concerning the place category, we generalised location (LOC) to place (PLC). This category includes geopolitical entities, aquifers, mountains, facilities, and one extra subcategory for other locations.

ORG category includes all typologies of organisations, like, for example:

- Convento de Santo António [Santo António Monastery]
- Santo Ofício de Évora [Tribunal of the Holly Office of Évora]
- Confraria de São Pedro [São Pedro Fraternity]

For Time, we only annotated specific reference to dates, for instance, o ano de 1755 [the year of 1755].

Our subcategories were chosen based on the fact that in the 18^{th} century, there was still inequality of each person before the law and hierarchy structured the Portuguese society. Frequently, titles and occupations positions were almost part of a person's name and identity. Also, the organisations had different societal roles, and the difference between a location and a geopolitical organisation may be thin. Other references to geographical points, such as rivers and mountains, are essential for geo-references. These were some of the reasons that supported the need to reestablish the NEs to describe the elements of the source better and to make the annotation process more relevant from the point of view of History. However, this is a challenging question. A more detailed and adequate establishment of NE categories to past ages frequently implicates more complexity in annotation and their computational processes, which we assumed from the beginning.

4.2 Annotation guidelines

As usual in this kind of study, annotation guidelines were defined as a basis for the manual annotation process. The construction of the guidelines was a vital phase in the manual annotation process, as there were several annotators, and all must have the same decision support. All categories and subcategories have examples from different *corpus* texts detailing different complex situations.

The delimitation should include the totality of the expression, including additional sequential information such as apposition. That decision was related to the importance of entity disambiguation. The two first examples show that the annotation of all the expression and not just the name is vital to disambiguate:

- Morgado Francisco José Cordovil - where "Morgado" is not part of the name but an identification for a holder of an entail estate
- Dom Frei João de Azevedo bispo - in this case we maintained Dom and Frei [Friar] as it is a mention of the statute, and they are both part of the name
- Francisco José Cordovil, natural de Évora - here we include the additional information natural de Évora [born in Évora]

In the guidelines, we also established that only NEs that include proper names should be annotated. For example, we should annotate the expression "cabido da Sé de Évora" [chapter of the Cathedral of Évora], but not the single uses of "cabido" [chapter]. In another example, we should mark the organisation "Santa Casa da Misericórdia de Beja", not just the general name "misericórdia".

4.3 Annotation process

All transcribed texts were manually normalised to standard European Portuguese to diminish spelling variance. The manual annotation was conducted over normalised texts and as a consensual process, with four annotators sharing the screen and deciding what to annotate. The annotators team comprised a linguist, two historians, and a computer scientist. During this process, the guidelines were reviewed when needed. After that initial phase of the definition of criteria and building of a consensual annotation, one historian proceeded with the task, bringing doubts to the team for discussion when they appeared.

The annotation tool used was the INCEPTION platform[1].

4.4 Annotated *corpus* description

The annotated subset gathers 71 parishes of Alentejo, corresponding to 17% of parishes of this region, the largest in Portugal. However, qualitatively, they belong to the most important municipalities: Beja, Évora, Portalegre and Vila Viçosa. The first three are the district capitals nowadays. Vila Viçosa, in the past, was the headquarters of the Duke of Bragança.

As we can see in Table 1, as a result of the manual annotation we have 5031 annotated NEs. The distribution is unbalanced, where the major categories represented in the corpus are related to geopolitical entities, person names, and saints. Persons referenced only by category and mountains are the less represented ones. Note that for the learning process, described in the sequence, they had to be separated for training, development and testing, considering approximately a distribution of 70, 10 and 20%.

CATEG	Train	Dev	Test	Overall NE
AUTWORK	106	12	19	137
ORG	287	52	54	393
PER_AUT	101	13	15	129
PER_CAT	37	4	8	49
PER_DIV	119	25	40	184
PER_NAM	520	62	136	718
PER_OCC	88	11	25	124
PER_PGRP	153	25	21	199
PER_SAINT	435	76	133	644
PLC_AQU	147	13	68	228
PLC_FAC	202	18	69	289
PLC_GPE	785	84	232	1101
PLC_LOC	336	24	87	447
PLC_MOUNT	50	10	13	73
TIM_CRON	217	33	66	316
Total	3583	462	986	5031

Table 1: Distribution of the quantity of Named Entities for the training, development, and test sets. The 'Overall NE' column represents the sum of the values from the three preceding columns.

5 Computational resources for building annotation models

5.1 Flair Framework

Flair(Akbik et al., 2019) is a NER library for multiple languages developed in PyTorch[2]. With Flair, we can construct pipelines for training token classifiers and feed them with various types of language models, such as Word Embeddings, Transformer-based models and Flair Embeddings itself. It is important to highlight that there are distinctions between the *Flair* framework and *Flair Embeddings* language models. *Flair Embeddings* are character-based models trained with recurrent neural networks, and the *Flair* library provides components for users to train models of this type.

Stacking Embeddings Combining language models for NER is beneficial, as demonstrated in the seminal Flair Embeddings article(Akbik et al., 2018). Within the *Flair* framework, we have a tool called *Stacking Embeddings* that allows the combination of different types of language models: transformer-based models, Flair embeddings, and shallow WE. Thus, each word is represented by

[1] https://inception-project.github.io

[2] https://pytorch.org/

the concatenation of vectors provided by each language model loaded into the *Stacking Embeddings*.

Sequence Tagger The introduction of the LSTM-CRF neural architecture for labelling token sequences was a milestone in the task of named entity recognition(Lample et al., 2016). With the advent of Transformer-based models like BERT, a new approach to entity recognition emerged. In this context, we adopted two types of structures for tagging the Parish Memories: the traditional LSTM-CRF and Transformer-Linear.

LSTM-CRF is essentially composed of two components: the Long-Short Term Memory (LSTM) neural structure(Hochreiter and Schmidhuber, 1997) and a Conditional Random Fields (CRF) classifier(Lafferty et al., 2001). First, an *embeddings* layer receives the *Stacked Embeddings* and then converts the input tokens into context-enriched vectors. Subsequently, these vectors are fed into the LSTM, which learns annotation patterns, and finally, the CRF classifier receives the outputs and returns the label sequence.

Transformer-Linear consists of a Transformer-based language model, to which a final linear layer is added to return the label sequence. This strategy aligns with the one applied in the seminal BERT article(Devlin et al., 2019). This fine-tuning approach is also available within the *Flair* framework and has been integrated into *Flair* as *Flert*(Schweter and Akbik, 2020). In this way, we also utilized Flair to train the model with Flert.

5.2 HappyTransformer

A less explored approach to sequence labelling is to use text-to-text algorithms. These algorithms take text as input and produce text as output. They are also known as sequence-to-sequence (Seq2Seq) algorithms. In this context, we used the HappyTransformer framework to train our Seq2Seq model for named entity recognition.

5.3 Embeddings

In this work, we used three types of Language Models: Shallow Word Embeddings, Contextual Embeddings, and Large Language Models. Below, we present the models used and their configurations.

Shallow Word Embeddings The use of Word Embeddings (WE) in the NER task dates back to the advent of these language models and is widely employed with recurrent neural networks. In this work, we utilized two types of pre-trained Word Embedding models: Word2Vec(Mikolov et al., 2013) (Skip-gram) and Glove(Pennington et al., 2014), both with 300 dimensions. These models are provided by the NILC embeddings repository[3].

Flair Embeddings As a Flair Embeddings type, we used the *FlairBBP* models[4] trained by (Santos et al., 2019). The authors trained the model with approximately four million tokens. Flair Embeddings are trained using a BiLSTM, where the model is trained to predict the next character in a sequence of tokens. Each *Flair Embeddings* model consists of two files: a *forward* model and a *backward* model. A linear operation combines the two models and provides a representation for each word, which is context-sensitive. This makes this type of model a contextual embedding, meaning that the representations change according to the context. This embedding type differs from Word Embeddings (WE), as WE uses fixed vectors. We experimented with *Flair Embeddings* models due to their unique versions for Portuguese and their ease of use.

XLM-R XLM-RoBERTa(Conneau et al., 2020) is a multilingual language model of the RoBERTa type. This model was pretrained on a 2.5 TB corpus of data containing one hundred languages. Out of the total of 2.5 TB training data, 49.1 GB consisted of Portuguese data, which amounts to approximately 8.4 billion tokens. We can describe XLM-RoBERTa by first describing the original RoBERTa model. RoBERTa is based on transformers and is pretrained on a large unsupervised corpus. RoBERTa inherits the masked language model training strategy from BERT, where the model's objective during training is to predict the masked tokens in a sentence. During the training phase, 15% of the input tokens were masked for prediction.

In this article, we used the Large version of XLM-R, which is available in the HuggingFace repository[5]. We chose this model type because it is extremely competitive with the current state of the art in English NER.

BERTimbau BERTimbau(Souza et al., 2020) is a BERT-style pretrained language model trained for Portuguese. This model was trained on the *brWaC* corpus(Filho et al., 2018), which amounts

[3] http://nilc.icmc.usp.br/nilc/index.php

[4] https://github.com/jneto04/ner-pt

[5] https://huggingface.co/xlm-roberta-large

to a total of 2.6 billion tokens, resulting in 17.5 GB of preprocessed data. We used the Large version of BERTimbau, which is available on HuggingFace[6].

BERTimbau is a transformer-based model and was also trained using token masking in input sentences. We chose this model because the current state-of-the-art(Souza et al., 2019) in NER for Portuguese utilizes this model.

LLaMa 2 We used two versions of LLaMa2(Touvron et al., 2023) through HuggingFace: the original version[7] (provided by Meta) and a version trained by NousResearch[8]. In both cases, we utilized the *chat* version with 7 billion parameters. The pretrained LLaMa 2 models were trained on 2 trillion tokens and fine-tuned with over 1 million human annotations.

The training of LLaMa 2-Chat begins with pretraining using a Transformer architecture on publicly available online data sources. Then, supervised fine-tuning is performed to create an initial version of *LLaMa 2-chat*. Finally, a refinement phase is initiated through an interactive process using Reinforcement Learning with Human Feedback (RLHF) methodologies.

5.4 Reduction tools

There are many advantages to using LLMs, but one of their disadvantages is the computational power required for their use, whether for inference or fine-tuning. It is in this context that we employed techniques for parameter reduction and model weight precision reduction. In this section, we define these techniques and how we apply them. These two techniques were used only on the two LLaMa models evaluated in this study.

Quantisation The quantisation technique comes from statistics, which is the process of mapping infinite continuous values into a finite discrete set. In the context of LLMs, the reduction occurs in the precision of the weights, which, in the case of LLaMa2, are initially 32 bits. In this regard, we converted our model to an 8-bit precision using the bitsandbytes library(Dettmers et al., 2022).

PEFT-LoRA After quantisation, we efficiently fine-tuned the model using PEFT-LoRA(Mangrulkar et al., 2022; Hu et al., 2022), where the authors demonstrated that freezing model weights and reducing the complexity of the matrices in the Transformer layers, significantly reduces the number of parameters while still yielding results equal to or better than the original model. In other words, PEFT-LoRA reduces the number of trainable parameters during fine-tuning. We used a rank $r = 64$ and $\alpha = 16$.

### Instruction: "Recognize named entities and rewrite each input token followed by its label until the end of the input sentence."
### Input: "Tem catorze moinhos , na Ribeira de Caia , e Caldeirão , e três pisões ."
### Response: "Tem <|O|> catorze <|O|> moinhos <|O|> , <|O|> na <|O|> Ribeira <|B-PLC_AQU|> de <|I-PLC_AQU|> Caia <|I-PLC_AQU|> , <|O|> e <|O|> Caldeirão <|B-PLC_AQU|> , <|O|> e <|O|> três <|O|> pisões <|O|> . <|O|>"

Figure 1: Instruction example

Input: "ner: Tem catorze moinhos , na Ribeira de Caia , e Caldeirão , e três pisões ."
Target: "Tem <|O|> catorze <|O|> moinhos <|O|> , <|O|> na <|O|> Ribeira <|B-PLC_AQU|> de <|I-PLC_AQU|> Caia <|I-PLC_AQU|> , <|O|> e <|O|> Caldeirão <|B-PLC_AQU|> , <|O|> e <|O|> três <|O|> pisões <|O|> . <|O|>"

Figure 2: Text-to-Text training example

5.5 Needleman-Wunsch algorithm

The Needleman-Wunsch algorithm(Needleman and Wunsch, 1970) is a dynamic programming algorithm designed to align two sequences. This algorithm is commonly used for aligning protein or nucleotide sequences. In this work, we employed this algorithm to align the text labelled by the LLaMa and mT5 models with the gold standard text, enabling the extraction of evaluation metrics. We used the implementation provided by Genalog[9] in Python.

6 Experiments

6.1 Experiments Configuration

We have two sets of experiments: (i) Experiments with LLMs and (ii) Experiments with stacking embeddings. Starting with the set of experiments

[6] https://huggingface.co/neuralmind/bert-large-portuguese-cased

[7] https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

[8] https://huggingface.co/NousResearch/Llama-2-7b-chat-hf

[9] https://microsoft.github.io/genalog/text_alignment.html

(i), we evaluated three LLMs: XLM-R, BERTimbau, LLaMa 2, and mT5. For the experiments conducted with XLM-R and BERTimbau, we used *Flert*, where the sequence tagging is composed by the model itself plus a final linear layer that returns the label sequence. Following the naming convention of (Schweter and Akbik, 2020), we refer to these experiments as *Transformer-Linear* since both evaluated models are based on transformers. We executed these experiments on one RTX 4090 GPU with 24GB of memory and used default hyperparameters.

Regarding the experiments with LLaMa 2, we performed *instruct-tuning*, where the prompt consists of an instruction, input, and response. Figure 1 shows an example of a prompt. To generate prompts, we created a script that reads the original CoNLL-formatted file and provides the sentence without annotations and another with annotations. For each example in the corpus, we added the same instruction: *Recognize the named entities and rewrite each input token followed by its label to the end of the input sentence.* We added three special tokens to the tokenizer: <*s*>, </*s*>, and <*unk*>, corresponding to *bos* (beginning of sentence), *eos* (end of sentence), and *pad* (padding). We defined the start-of-sentence token to be the first token of the prompt and the end-of-sentence token to be the last. It is essential to define the end of the sentence with a special token to ensure that the model learns to stop generating text, thus preventing hallucinations. In the tokenizer, we set an input size of 1024 tokens, and during prediction, we defined a maximum of 512 new tokens. Once the instruction corpus was ready, we performed instruct-tuning using the HuggingFace training pipeline for Causal models. To reduce computational costs, we employed the *Quantization* technique, which converts the model to an 8-bit precision. We also used PEFT-LoRa, which reduces the number of trainable parameters. With these reductions, we were able to carry out fine-tuning on a Tesla T4 GPU with 16GB.

Regarding the experiment conducted with mT5, we used a *Text-to-Text* algorithm pipeline provided by the HappyTransformer framework[10]. Only the input and output sizes were modified to 512 tokens, while the other hyperparameters remained the same. Similar to what we did to prepare the data for LLaMa2's *instruct-tuning*, we created a script that returns two types of sentences from the original CoNLL data. The algorithm generates input sentences (containing only text without annotations) and target sentences (containing tokens followed by their respective labels). Figure 2 shows an example. Therefore, the *Seq2Seq* algorithm takes the sentence without named entities and is trained to generate a sentence with identified and classified entities. Note that the input sentence receives a *ner:* prefix to indicate that the task it is learning is entity recognition. We conducted this experiment on an RTX 4090 24GB GPU.

For the set of experiments (ii), we used the Vanilla LSTM-CRF implemented in *Flair*. Thus, we created two stack embeddings: *FlairBBP + Word2Vec (Skip-gram*, hereinafter referred to as *FlairBBP+W2V-SKPG*, and *FlairBBP + Glove*. We combined these embeddings because (Santos et al., 2019) showed that combining FlairBBP with Word2Vec (skip-gram) was the best stack embedding for named entity recognition in the HAREM corpus(Santos and Cardoso, 2007). In the original work on *Flair*, the authors stacked a *Flair Embeddings* model with a Glove language model. However, this experiment was not conducted by (Santos et al., 2019). Therefore, we decided to evaluate this stack embeddings. We executed both experiments on an RTX 4090 24GB GPU.

6.2 Evaluation and Metrics

The models trained using the *Transformer-Linear* approach and the vanilla LSTM-CRF were directly evaluated using the named entity recognition evaluation script from CoNLL-2002(Sang and Erik, 2002). We chose this script because it is commonly used in NER research for both Portuguese and English. The script returns the Precision (PRE), Recall (REC), and F_1 metrics for each category and for the entire predicted corpus.

The evaluation of the mT5 and LLaMa2 models requires preprocessing before being evaluated by the script. The preprocessing consists of:

- Aligning the key sentences with the sentences predicted by the model. This alignment is performed using the Needleman-Wunsch algorithm.
- Separating punctuation that is combined with tokens. This was a common issue in mT5 predictions.
- Sometimes labels may contain the symbol @

[10] https://github.com/EricFillion/happy-transformer

Architecture	Model	PRE	REC	F_1	$\Delta\uparrow$	$\Delta\downarrow$
Transformer-Linear	**XLM-R-Large**	**68.31**	73.38	**70.76**	+0.23	*sota*
	BERTimbau-Large	67.36	**74.00**	70.53	+3.03	−0.23
LSTM-CRF	**FlairBBP + W2V-SKPG**	67.77	67.23	67.50	+1.23	−3.03
	FlairBBP + Glove	66.50	66.04	66.27	+17.24	−1.23
Causal LM	**LLaMa 2 (8bit) + LoRa**	68.01	38.34	49.03	+6.28	−17.24
Text-to-Text	**mT5-Large**	48.55	38.19	42.75	*bl*	−6.28

Table 2: Overall metrics. *bl* = baseline and *sota* = state-of-the-art.

CATEG	XLM-R			BERTimbau			LlaMa 2			mT5		
	PRE	REC	F_1	PRE	REC	F_1	PRE	REC	F_1	PRE	REC	F_1
AUTWORK	47.83	55.00	51.16	45.83	52.38	48.89	100.00	6.25	11.76	100.00	5.56	10.53
ORG	53.23	55.93	54.55	48.05	67.27	56.06	23.53	09.09	13.11	28.00	23.33	25.45
PER_AUT	78.95	93.75	85.71	77.78	87.50	82.35	100.00	50.00	66.67	0.00	0.00	0.00
PER_CAT	50.00	75.00	60.00	87.50	87.50	87.50	57.14	57.14	57.14	0.00	0.00	0.00
PER_DIV	69.57	80.00	74.42	76.74	82.50	79.52	88.24	38.46	53.57	57.14	23.53	33.33
PER_NAM	66.23	71.83	68.92	61.04	67.63	64.16	49.46	34.07	40.35	44.44	53.12	48.40
PER_OCC	60.71	62.96	61.82	44.12	60.00	50.85	66.67	09.09	16.00	50.00	4.00	7.41
PER_PGRP	55.17	76.19	64.00	50.00	61.90	55.32	100.00	5.26	10.00	0.00	0.00	0.00
PER_SAINT	75.69	78.99	77.30	77.37	79.10	78.23	87.34	55.65	67.98	81.74	70.15	75.50
PLC_AQU	72.73	76.71	74.67	66.20	67.14	66.67	77.42	38.10	51.06	0.00	0.00	0.00
PLC_FAC	59.52	66.67	62.89	65.33	67.12	66.22	59.46	32.84	42.31	0.00	0.00	0.00
PLC_GPE	78.84	77.87	78.35	77.87	81.55	79.66	64.12	49.55	55.90	43.41	63.88	51.69
PLC_LOC	60.00	72.53	65.67	65.35	74.16	69.47	81.25	30.59	44.44	23.44	17.24	19.87
PLC_MOUNT	75.00	92.31	82.76	56.25	69.23	62.07	100.00	75.00	85.71	0.00	0.00	0.00
TIM_CRON	66.67	65.71	66.19	69.33	77.61	73.24	90.00	28.57	43.37	80.00	18.46	30.00

Table 3: LLMs results by category

due to the alignment phase. So, we replace the labels from the aligned sentence with the labels from the predicted sentence.

- Rewriting the sentences in CoNLL format. We do not include it in the final evaluation file the lines where the key token or label or the predicted label contains the symbol @.

We based our preprocessing pipeline on NER evaluation from generative models on (Paolini et al., 2021). Once preprocessing was completed, we applied the CoNLL-2002 evaluation script to obtain the metrics.

7 Experiment Results

In this section, we present our results. Table 2 shows the overall metrics for each evaluated model. From these general results, we establish the best and least favourable models for named entity recognition in our *Parish Memories corpus*. Two models had $F_1 > 70\%$ with a small difference between them, as shown in the columns $\Delta\uparrow$ and $\Delta\downarrow$.

CATEG	FlairBBP+W2V-SKPG			FlairBBP+Glove		
	PRE	REC	F1	PRE	REC	F1
AUTWORK	47.37	42.86	45.00	52.63	47.62	50.00
ORG	50.00	60.00	54.55	53.23	60.00	56.41
PER_AUT	81.25	81.25	81.25	70.59	75.00	72.73
PER_CAT	57.14	100.00	72.73	40.00	75.00	52.17
PER_DIV	78.95	75.00	76.92	73.17	75.00	74.07
PER_NAM	64.54	65.47	65.00	60.40	64.75	62.50
PER_OCC	88.24	60.00	71.43	75.00	60.00	66.67
PER_PGRP	43.75	66.67	52.83	41.38	57.14	48.00
PER_SAINT	73.57	76.87	75.18	71.64	71.64	71.64
PLC_AQU	75.00	60.00	66.67	72.73	57.14	64.00
PLC_FAC	63.46	45.21	52.80	54.55	41.10	46.88
PLC_GPE	71.77	76.39	74.01	73.64	75.54	74.58
PLC_LOC	61.45	57.30	59.30	64.71	61.80	63.22
PLC_MOUNT	63.16	92.31	75.00	80.00	92.31	85.71
TIM_CRON	78.18	64.18	70.49	74.19	68.66	71.32

Table 4: Vanilla LSTM-CRF results

Analysing the Precision metric (PRE), the *XLM-R-Large* model had the highest metric, meaning it was the best model for correctly identifying entities. On the other hand, the *BERTimbau-Large* model stood out in the Recall metric, indicating that it achieved the highest percentage of named entities found. When it comes to the F_1 metric, which combines both Precision and Recall, *XLM-R-Large* was the best-performing model. Regarding the use of *Glove*, continuing to use *W2V-SKP* is the better option.

From the perspective of the two generative models (LLaMa2 and mT5), we only present the metrics of the LLaMa2 model from NousResearch, as it showed considerably better performance compared to the original Meta model. Our evaluation reveals that the LLaMa 2 (original) model achieved an F_1 score of 42.71, a decrease of 6.32 points compared to the unofficial LLaMa's F_1. These LLMs had significantly lower results than the other models. We believe this is due to the limited number of examples available at the moment for some categories and the inherent complexity of certain categories. We base this hypothesis on the work of (Paolini et al., 2021), which showed competitive results in various sequence labelling tasks but with a much larger amount of training data. Therefore, based on the F_1 metric, we can conclude that the *XLM-R-Large* model was the best model.

CATEG	**Max**		**Min**	
	Model	F_1	**Model**	F_1
AUTWORK	XLM-R	51,16	mT5	10,53
ORG	Glove	56,41	LLaMa2	13,11
PER_AUT	XLM-R	85,71	LLaMa2	66,67
PER_CAT	BERTimbau	87,50	Glove	52,17
PER_DIV	BERTimbau	79,52	mT5	33,33
PER_NAM	XLM-R	68,92	LLaMa2	40,35
PER_OCC	W2V-SKPG	71,43	mT5	7,41
PER_PGRP	XLM-R	64,00	LLaMa2	10,00
PER_SAINT	BERTimbau	78,23	mT5	67,98
PLC_AQU	XLM-R	74,67	LLaMa2	51,06
PLC_FAC	XLM-R	62,89	LLaMa2	42,31
PLC_GPE	BERTimbau	79,66	mT5	51,69
PLC_LOC	BERTimbau	69,47	mT5	19,87
PLC_MOUNT	Glove	85,71	BERTimbau	62,07
TIM_CRON	BERTimbau	73,24	mT5	30,00

Table 5: Best and worst models by category.

Tables 3 and 4 present the comprehensive results for LLMs and LSTM-CRF, respectively, for each category in the corpus. We summarized these two tables into a smaller set, table 5. This table shows the model that achieved the maximum F_1 score for each category and also indicates which model had the lowest F_1 score (above 0%) for each category. We can observe that the *XLM-R* and *BERTimbau* models tied when referring to the number of maximum F_1 scores per category, followed by the stack embeddings with the Glove and W2V-SKPG models. This analysis allowed us to identify that the embeddings stack with Glove had better overall metrics than the stack containing W2V-SKPG, although the W2V-SKPG model remained more stable.

Regarding the minimums, mT5 had the highest number of minimum scores above zero, followed by LLaMa2. As seen in Table 2, mT5 also had the highest number of zeros. Note also that the stack containing Glove had the worst score above zero in the **PER_CAT** category, while *FlairBBP+W2V-SKP* was not the worst in any category. We also highlight that *BERTimbau* performed the worst in the **TIM_CRON** category.

Thus, we can see, after the experiments, that it is still much more advantageous to use a BERT-style model with a linear layer.

8 Conclusion

In this work, we present a *corpus* study for the task of named entity recognition based on 18^{th}century texts, produced by Alentejo parish priests, Portugal. For this study, motivated by the historians' research objectives, new NE categories were defined. As there were no previous models trained with these new categories, it was necessary to train new models. In this process, we evaluated several language models and architectures and our best model was *XLM-R-Large*, which can be trained on a single GPU, without the need for parameter reduction techniques and in just a few hours. Our evaluations involved multilingual and Portuguese-specific models, with only a small margin of difference in the metrics of the two best models, which are multilingual and monolingual (for Portuguese), respectively. With the current results, we believe it will be possible to use the models in an assisted-based annotation system to accelerate the annotation process of the whole collection of the *Parish Memories*.

In future work, we plan to refine models for 18^{th} century Portuguese and expand the *corpus* annotation.

Acknowledgements

This work has received financial support from the Brazilian funding agency CAPES and from the Portuguese Science Foundation FCT, in the context of the projects CEECIND/01997/2017 and UIDB/00057/2020.

References

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Hidelberg O Albuquerque, Ellen Souza, Carlos Gomes, Matheus Henrique de C Pinto, PS Ricardo Filho, Rosimeire Costa, Vinícius Teixeira de M Lopes, Nádia FF da Silva, André CPLF de Carvalho, and Adriano LI Oliveira. 2023. Named entity recognition: a survey for the portuguese language. *Procesamiento del Lenguaje Natural*, 70:171–185.

Helena Freire Cameron, Fernanda Olival, Renata Vieira, and Joaquim Francisco Santos Neto. 2022. Named entity annotation of an 18th century transcribed corpus: problems, challenges. In *Proceedings of the Second Workshop on Digital Humanities and Natural Language Processing (2nd DHandNLP 2022) co-located with International Conference on the Computational Processing of Portuguese (PROPOR 2022), Virtual Event, Fortaleza, Brazil, 21st March, 2022*, volume 3128 of *CEUR Workshop Proceedings*, pages 18–25. CEUR-WS.org.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, pages 8440–8451. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *CoRR*, abs/2208.07339.

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

Maud Ehrmann, Ahmed Hamdi, Elvys Linhares Pontes, Matteo Romanello, and Antoine Doucet. 2023. Named entity recognition and classification in historical documents: A survey. *ACM Comput. Surv.*, 56(2).

Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brwac corpus: a new open resource for brazilian portuguese. In *Proceedings of the 11th International conference on language resources and evaluation*, pages 4339–4344.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations*. OpenReview.net.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International conference on machine learning*, pages 282–289.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the Conference of the north american chapter of the association for computational linguistics: human language technologies*, pages 260–270.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the 1st International Conference on Learning Representations*.

Saul B. Needleman and Christian D. Wunsch. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453.

Giovanni Paolini, Ben Athiwaratkun, Jason Krone, Jie Ma, Alessandro Achille, Rishita Anubhai, Cícero Nogueira dos Santos, Bing Xiang, and Stefano Soatto. 2021. Structured prediction as translation between augmented natural languages. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics.

Tjong Kim Sang and F Erik. 2002. Introduction to the conll-2002 shared task: language-independent named entity recognition. In *Proceedings of CoNLL-2002*, pages 155–158.

Diana Santos and Nuno Cardoso. 2007. Reconhecimento de entidades mencionadas em português: Documentação e actas do harem, a primeira avaliação conjunta na área.

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

Stefan Schweter and Alan Akbik. 2020. FLERT: Document-level features for named entity recognition.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems, BRACIS*.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

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