Journal Title
XX(X):1–8

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Ontology-based Information Extraction from Cultural Heritage Digital Representations: A Case Study in Portuguese Archives

Mariana Dias[1] and Carla Teixeira Lopes[1]

Abstract

Linked Data (LD) enables cultural heritage institutions to refine archival descriptions and improve findability, but manually creating LD descriptions remains labor-intensive. This paper presents an ontology-guided information extraction system that assists archivists by automatically identifying concepts and relations in digitized archival records. Focusing on Portuguese archival collections, we extract and structure data according to ArchOnto, a CIDOC-CRM-based LD model for archives, to support future metadata enrichment. Our approach identifies core archival entities from textual digital representations of archival records obtained through optical character recognition and human-made transcriptions. However, it shows limited results in extracting some entities and relational facts. Our low-performing results indicate that fine-tuning information extraction models using adapted general-domain datasets for Cultural Heritage tasks in 20th-century documents is only marginally viable.

Keywords

Information extraction, Cultural heritage, Linked data, Archives

Introduction

Linked Data (LD) emerged as a way to structure and connect data, enabling interoperability across various domains, including cultural heritage (1). LD can enhance archival descriptions by improving metadata quality through more accurate contentwhile broadening access to culturally enriched archives. This can give users deeper and more comprehensive insights into collections, enriching their understanding and engagement with cultural resources. However, manually creating LD descriptions for cultural heritage objects can be taxing and time-consuming, making describing documents or collections in finer detail challenging. Automating the extraction of relevant information from digitized archival records to create LD descriptions can alleviate this burden for cultural heritage professionals.

Our objective is to extract concepts and relationships from Portuguese digitized archival records and map them to ArchOnto (2), an LD model built on CIDOC CRM (Conceptual Reference Model) (3) for archival description (4). To achieve this, we implemented an ontology-based information extraction (OBIE) system guided by ArchOnto. We trained information extraction models on contemporary Portuguese datasets and evaluate their performance in the Cultural Heritage domain using two datasets of 20th century Portuguese historical archival records: one containing text obtained through Optical Character Recognition (OCR) and the other with human-made transcriptions.

We aim to address the following research question: how effective is fine-tuning information extraction models with general-domain annotated datasets for extracting Cultural Heritage-specific entities and relations from 20th-century texts, including OCR-extracted text and human-made transcriptions?

The trained models and datasets used in this work are privately shared* and will later be published in an open-access research data repository.

Background and Related Work

Information extraction (IE) involves identifying pre-specified types of information, such as entities, relations, and events, from unstructured or semi-structured information and structuring them (5). OBIE refines IE by using ontologies to define domain-specific concepts and properties, guiding the extraction process (6).

NER in Portuguese Cultural Heritage

Recent advances in machine-learning-based Named Entity Recognition (NER) for Portuguese Cultural Heritage have been driven by neural architectures. Table 1 summarizes key studies, detailing their approaches, annotated corpora, NER datasets outcomes, and evaluation results.

Vieira et al. (7) and Zilio et al. (11) employed models trained on contemporary Portuguese corpora but applied them to historical 18th century text, while Cunha et al. (15)

[1]Faculty of Engineering, University of Porto, Porto, Portugal

Corresponding author:
Mariana Dias
Email: up201606486@fe.up.pt

*https://figshare.com/s/cde1ccdfffbae587945d

Table 1. Comparison of NER research approaches in Portuguese Cultural Heritage.

Approach	Testing dataset	Training dataset	Model	Evaluation
Vieira et al. (7)	Parish Memories (1758–1761) (8)	First HAREM (9)*	BiLSTM-CRF + FlairBBP	F1=45.8%
		WikiNER (10)*	CNN	F1=38.4%
Zilio et al. (11)	18th-century medical texts (12; 13; 14)	First HAREM (9)*	BERT-CRF	FM=83.7%
		WikiNER (10)*	CNN (spaCy_lg)	FM=55.5%
			CNN (spaCy_sm)	FM=63.7%
Cunha et al. (15)	Dataset from Portuguese Archives (16) withheld from training	Datasets from Portuguese Archives (16; 17)	BiLSTM-CRF	F1=53.0%
			CNN	F1=68.4%
			Maximum Entropy	F1=66.6%
Santos et al. (18)	Manually annotated subset of the Parish Memories dataset		BERTimbau-Large	F1=70.5%
			BiLSTM-CRF + FlairBBP + W2V-SKPG	F1=67.5%
			BiLSTM-CRF + FlairBBP + Glove	F1=66.3%
			LLama2 (8bit) + LoRa	F1=49.0%
			mT5-large	F1=42.8%
			XLM-R-Large	F1=70.8%

*Contemporary dataset
F1 = F1-score; FM = rate of full matches

and Santos et al. (18) trained their models directly on digitized archival data from the 18th century.

Although the identified studies used similar architectures, performance varied significantly across datasets. For instance, CNNs achieved F1-scores ranging from 38.4% in Vieira et al. (7) to 68.4% in Cunha et al. (15). Moreover, BiLSTM-CRF performed best in Vieira et al. (7) and had optimal results in Santos et al. (18), but underperformed in Cunha et al. (15). This performance difference may be due to Cunha et al.'s BiLSTM-CRF architecture not leveraging pre-trained embeddings, unlike Vieira et al. and Santos et al., who employed Flair embeddings (FlairBBP) with their BiLSTM-CRF implementation.

Zilio et al. (11) and Santos et al. (18) evaluated transformer architectures with LLama2, mT5, and masked language models, such as BERT-based models and XLM-R for NER in Portuguese digitized archives. The results of their work report that BERT-based and XLM-R models achieved the best results. In contrast, LLama2 and mT5 yielded F1-scores below 50%.

The approaches also varied in entity scope, although all included person, location, and organization entities. Cunha et al. (15) added date and profession entities, while Zilio et al. (11) included time, work, event, value, and other entities, and Santos et al. (18) included time and author work entities.

RE in Cultural Heritage

Relation Extraction (RE) enable the identification and categorization of meaningful relationships and associations between cultural artifacts, historical events, and other entities within the domain of cultural heritage. Efremova et al. (19) compared the performance of a Support Vector Machine classification approach to a Hidden Markov Model to extract family relationships in historical notary acts. Chantaraj et al. (20) implemented a rule-based approach to extract relationships in Thai Buddhist temple documents. Christou et al. (21) employed deep-learning to extract semantic relationships in 19th-century Greek Literature documents.

To our knowledge, RE has not been explored in the domain of Portuguese cultural heritage. However, some works have applied RE to the broader domain of the Portuguese language (22; 23). Portuguese evaluation contests, such as HAREM (24) and IberLEF 2019 (25), included RE tasks. Santos et al. (22) created a rule-based system to extract family relations. Collovini et al. (23) evaluated a CRF model for open-relation extraction using the HAREM corpora and reported a 26% F1-score increase in extracting the placement relation compared to other RE systems.

Ontology-based Named Entity Recognition

We trained and evaluated BiLSTM-CRF models with Flair Embedding (26) models for Portuguese archival NER, as related work confirms that this architecture performs well whether trained on contemporary or archival datasets. This approach balances accuracy, computational efficiency, and robustness to noisy data (27), such as OCR errors and cartographic variations.

Creation of train and test datasets

Given the scarcity of manually curated annotated archival collections in Portuguese, we adapted contemporary general-domain NER corpora to align with the ArchOnto ontology's classes. We adapted the HAREM Golden Collections (GC), which consist of two datasets: the First HAREM GC (9) and the Second HAREM GC (24). The First HAREM GC includes 1,202 documents and 5,132 named entities, while the Second HAREM GC contains 1,040 documents, 7,847 named entities, and 4,803 relations. The collections cover a range of entity categories, such as Person, Location, Organization, Time, Value, Abstraction, Event, Thing, Work, and Other, along with 38 relation labels, both symmetric and direct/indirect, such as identity (*ident*), inclusion (*inclui/incluido*), and location (*ocorre_em/sede_de*). The pipeline for the creation of the training and test datasets, Arch.NER.Train and Arch.NER.Test, adapted from the HAREM GCs, is shown in Figure 1.

First, we aligned ArchOnto concepts with the HAREM GCs' classification labels by selecting named entity labels that describe entities representable with ArchOnto. We detail the mapping between HAREM GCs concepts and ArchOnto

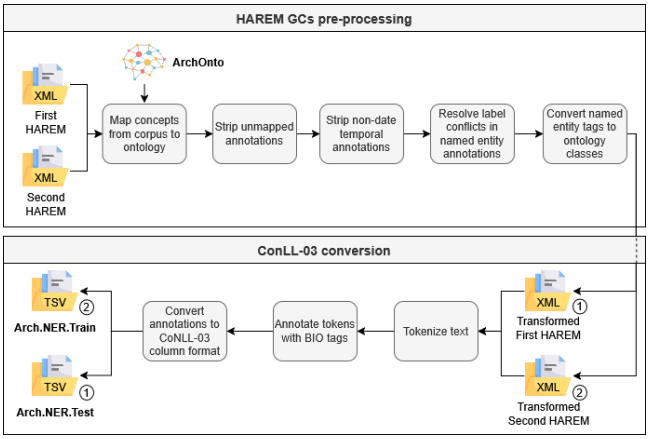

Figure 1. Pipeline for creating training and testing datasets for the NER task.

classes based on the CIDOC CRM core and extension in Table 2.

We then excluded unmapped and non-date temporal annotations, such as "passada sexta-feira" (last Friday), "mais tarde" (later), "ontem" (yesterday), and "há quinze anos" (fifteen years ago), as they are incompatible with ArchOnto's structure.

The GCs contained ambiguous entities with overlapping classifications that needed to be disambiguated (e.g., "Twin Towers" was classified as E5 Event and E53 Place). We manually resolved these conflicts using context clues while ensuring consistency among related entities. For example, if an ambiguous entity shared inclusion relationships with other named entities, such as a place nested within a place and a group within a group, all were classified under the same ontology class to ensure hierarchical coherence.

Subsequently, HAREM tags were systematically converted to ArchOnto labels based on the previously described alignment, resulting in the Transformed First and Second HAREM datasets.

Following the CoNLL-03 (28) (Computational Natural Language Learning) standard, we converted the transformed HAREM corpora by tokenizing XML content and implementing BIO tagging (Beginning, Inside, and Outside). The resulting datasets, Arch.NER.Train (Second HAREM) and Arch.NER.Test (First HAREM), are formatted as shown in Table 3, exemplified by the excerpt "Hugo Doménech, professor at the University Jaume de Castellón".

Evaluation datasets We further evaluate the NER models using two annotated datasets derived from 13 Portuguese 20th-century archival records extracted from the Torre do Tombo National Archive: one containing OCR extracted text (Arch.NER.Eval.OCR), and the other with manually transcribed text (Arch.NER.Eval.Human). These two datasets are subsets of publicly available datasets described in prior work (29). Both datasets were annotated according to consensual descriptions (30) developed by two archivists from the General Directorate for Book, Archives and Libraries (DGLAB) using the ArchOnto ontology, based solely on the textual content visible in the digital representations. Only the digital representations used by the archivists for the archival record descriptions are included in the datasets.

The distribution of named entities in the Arch.NER datasets is shown in Table 4, revealing class imbalances between entity types.

It should be noted that Arch.NER.Eval.OCR contains fewer entities than Arch.NER.Eval.Human due to OCR errors exacerbated by the lower quality of some documents and the inclusion of handwritten text.

Training of Machine Learning models

We used Flair NLP[†] (31), an open-source library for sequence labeling, to train NER models for Portuguese archival entities. We employed two pre-trained contextual language models: FlairBBP[‡] (32), bidirectional Flair embeddings trained on 4.9B tokens from three large Portuguese corpora, and FlairEL[§] (33), Flair embeddings trained on 0.9B tokens of Portuguese CommonCrawl data. We also used a pre-trained word embedding model, Skip-gram Word2Vec[¶]. We adopted the training hyperparameters set by Santos et al. (32) to avoid the computational cost of optimizing hyperparameters. We set the initial learning rate to 0.1 with an annealing factor of 0.5 on every three epochs without improvement. The training was stopped after 150 epochs or if the learning rate was below 0.0001.

Evaluation of the models

Following SemEval 2013 (34) guidelines, we evaluated models using strict matching and type matching. Strict matching requires exact named entity matches (35), while type matching considers the correctness of named entity terms regardless of boundaries.

Contemporary general-domain dataset As previously stated, we evaluated the models on a contemporary dataset, Arch.NER.Test, to establish a baseline for comparison with archival data. Table 5 presents F1-scores for the BiLSTM-CRF NER models, organized by entity label, embedding models, and evaluation scenario in the Arch.NER.Test dataset.

In the strict matching evaluation scenario, models that employed a Skip-gram Word2Vec word embedding model generally outperformed others on nearly all named entity categories, with the BBP+W2V combination achieving the highest overall score. As expected, the most frequently annotated entities, E21 Person, E52 Time-Span, E53 Place, and E74 Group, achieved the best results. However, the models struggled to recognize events, titles, and roles, achieving F1-scores below 40%.

Type matching substantially improved recognition for E52 Time-Span and E54 Dimension by resolving boundary-related errors. For events, titles, and roles, however, the difference between type matching and strict matching is less apparent, indicating that the models have more difficulty

[†] https://github.com/flairNLP/flair

[‡] Repository available online at https://github.com/jneto04/ner-pt

[§] Repository available online at https://github.com/ericlief/language_models

[¶] Word2Vec Skip-gram 300 dimensions available at http://nilc.icmc.usp.br/nilc/index.php/repositorio-de-word-embeddings-do-nilc

Table 2. Mapping between concepts in the HAREM GCs and ArchOnto classes.

Concept	HAREM category	HAREM type	HAREM subtype	ArchOnto class
Title assigned to works	Obra (Work)	Arte (Art)	Pintura (Painting)	ARE2 Formal Title
		Plano (Plan)	-	
		Reproduzida (Reproduced)	Livro (Book), Programa (Program), Musica (Music), Outro (Other)	
Role of a person in an event	Pessoa (Person)	Cargo (Position), GrupoCargo (GroupRole)	-	ARE8 Role Type
Event	Acontecimento (Event)	Efemeride (Anniversary), Evento (Event), Organizado (Organized), Outro (Other)	-	E5 Event
Person	Pessoa (Person)	Individual, Membro (Member)	-	E21 Person
Temporal expression	Tempo (Time)	Data (Date), Periodo* (Period)	-	E52 Time-Span
		Tempo_Calend** (Time_Calend)	Data (Date), Intervalo (Interval)	
Location	Local (Place)	Humano (Human)	Rua (Street), Pais (Country), Divisao (Division), Regiao (Region), Outro (Other)	E53 Place
		Fisico (Physical)	Ilha (Island), Aguacurso (Water course), Planeta (Planet), Regiao (Region), Relevo (Relief), Aguamassa (Water mass), Outro (Other)	
		Outro (Other)	-	
Quantity	Valor (Value)	Quantidade (Quantity)	-	E54 Dimension
Organization	Organizacao (Organization)	Administracao (Administration), Empresa (Company), Instituicao (Institution), Outro (Other)	-	E74 Group
	Pessoa	GrupoInd (GroupInd), GrupoMembro (GroupMember)	-	

* Mapping with First HAREM GC; ** Mapping with Second HAREM GC

Table 3. Example of converting an annotated textual segment from HAREM to the ConLL-03 format of Arch.NER.Test and Arch.NER.Train.

Natural Language	Hugo Doménech, professor da Universidade Jaume de Castellón
XML	<E21>Hugo Doménech</E21>, <ARE8>professor</ARE8> da <E74>Universidade Jaume de Castellón</E74>
CoNLL-03	Hugo B-E21 Doménech I-E21 , O professor B-ARE8 da O Universidade B-E74 Jaume I-E74 de I-E74 Castellón I-E74

Table 4. Distribution of recognized named entities in Arch.NER datasets.

Entities	Train	Test	Eval.Human	Eval.OCR
ARE2 Formal Title	503	267	7	5
ARE8 Role Type	178	92	26	19
E5 Event	313	156	1	0
E21 Person	1,863	1,333	12	8
E52 Time-Span	644	462	19	12
E53 Place	1,390	1,639	13	10
E54 Dimension	247	666	0	0
E74 Group	1,591	1,281	27	19
Total	6,730	5,898	105	73

in identifying the correct entity types, rather than simply making boundary errors.

Twentieth-century archival dataset We evaluated the NER models on Arch.NER.Eval.OCR, a dataset of Portuguese 20th-century digitized archival records with content extracted via OCR, as well as on Arch.NER.Eval.Human, which contains the same documents but with human-made transcriptions. Tables 6 and 7 present the performance of the NER models on the OCR-extracted and human-transcribed versions of the dataset, respectively.

The NER models demonstrated higher performance on the Arch.NER.Eval.Human dataset rather than on Arch.NER.Eval.OCR, as expected, given the negative impact of OCR errors on text quality. Despite this, strict matching performance remains low, particularly for titles, events, and groups.

The improved results in the type matching evaluation compared to the strict evaluation emphasize how the NER models are more successful at identifying the correct type of entity than at accurately determining entity boundaries.

No single model or embedding consistently outperforms others across all entity types, which indicates that the embeddings' effectiveness depends on the specific entity type and the quality of the input text, rather than solely on model size.

Ontology-based Relation Extraction

We trained and tested CRF classifiers for three types of relations: affiliation (between Person and Group), authorship (between Title and Person), and placement (between Group and Place).

Creation of datasets

The Second HAREM contest produced the ReRelEM (Recognition of Relations between Mentioned Entities) GC, a Portuguese relation-annotated dataset. We used the ReRelEM dataset to develop training and evaluation datasets for RE models, as outlined in Figure 2.

We initially mapped ReRelEM GC's labels to ArchOnto concepts and relations, identifying 14 relevant relationships. Following the removal of unmapped annotations, solving of ambiguous annotations, and conversion of HAREM tags to ArchOnto labels, the transformed dataset displayed a highly

Table 5. F1-score performance of BiLSTM-CRF models using different embedding models on the Arch.NER.Test dataset.

Entities	Strict Match				Type Match			
	EL	EL+W2V	BBP	BBP+W2V	EL	EL+W2V	BBP	BBP+W2V
ARE2 Formal Title	31.83%	**35.15%**	30.06%	33.12%	**47.94%**	44.85%	40.08%	45.86%
ARE8 Role Type	26.51%	30.84%	**31.62%**	29.79%	35.34%	41.41%	**45.30%**	38.30%
E5 Event	34.94%	**38.71%**	37.08%	36.88%	46.39%	**52.20%**	45.59%	44.38%
E21 Person	75.84%	77.36%	79.51%	**80.84%**	88.03%	90.64%	89.52%	**92.19%**
E52 Time-Span	**77.48%**	75.19%	75.49%	74.47%	91.22%	**92.17%**	90.02%	89.86%
E53 Place	84.65%	**86.49%**	85.00%	85.44%	87.40%	**88.81%**	87.10%	87.12%
E54 Dimension	**53.83%**	49.20%	45.43%	51.33%	**75.72%**	75.48%	71.22%	68.61%
E74 Group	66.94%	69.00%	68.33%	**71.31%**	73.05%	75.01%	73.79%	**76.57%**
Overall	70.10%	71.24%	70.48%	**72.47%**	79.50%	**81.27%**	79.37%	80.53%

Bold values indicate the best F1-score per line.

Table 6. F1-score performance of BiLSTM-CRF models using different embedding models on the Arch.NER.Eval.OCR dataset.

Entities	Arch.NER.Eval.OCR							
	Strict Match				Type Match			
	EL	EL+W2V	BBP	BBP+W2V	EL	EL+W2V	BBP	BBP+W2V
ARE2 Formal Title	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	**22.22%**	20.00%
ARE8 Role Type	0.00%	0.00%	0.00%	0.00%	**6.06%**	5.41%	0.00%	0.00%
E21 Person	11.76%	10.00%	0.00%	**26.67%**	11.76%	10.00%	13.33%	**26.67%**
E52 Time-Span	9.09%	9.09%	9.09%	**10.00%**	18.18%	18.18%	**27.27%**	20.00%
E53 Place	0.00%	**10.00%**	0.00%	0.00%	9.52%	**10.00%**	0.00%	0.00%
E74 Group	**3.51%**	3.08%	4.88%	4.44%	**17.54%**	15.38%	14.63%	13.33%
Overall	3.82%	4.65%	3.03%	**5.84%**	**12.74%**	11.63%	12.12%	11.68%

Table 7. F1-score performance of BiLSTM-CRF models using different embedding models on the Arch.NER.Eval.Human dataset.

Entities	Arch.NER.Eval.Human							
	Strict Match				Type Match			
	EL	EL+W2V	BBP	BBP+W2V	EL	EL+W2V	BBP	BBP+W2V
ARE2 Formal Title	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
ARE8 Role Type	**12.00%**	8.33%	4.88%	0.00%	**32.00%**	29.17%	24.39%	18.60%
E5 Event	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
E21 Person	5.26%	10.81%	**14.29%**	0.00%	31.58%	37.84%	**50.00%**	42.42%
E52 Time-Span	6.25%	**12.12%**	11.76%	12.90%	43.75%	54.55%	**58.82%**	58.06%
E53 Place	0.00%	**8.00%**	0.00%	0.00%	**41.67%**	32.00%	33.33%	16.00%
E74 Group	2.82%	2.70%	2.90%	**3.28%**	45.07%	51.35%	52.17%	**59.02%**
Overall	5.22%	**6.84%**	5.66%	2.88%	36.52%	39.32%	**41.51%**	38.46%

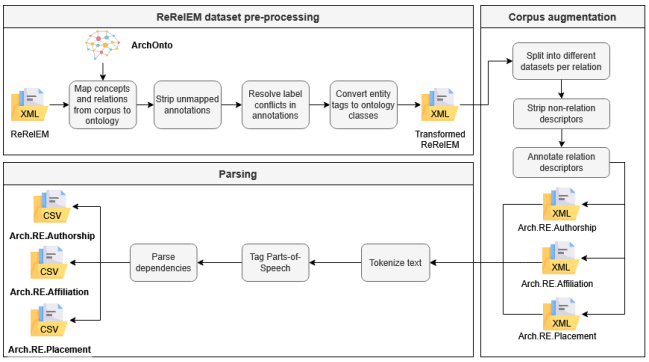

Figure 2. Pipeline for creating training and testing datasets for the RE task.

imbalanced distribution of relations between NEs, as can be seen in Table 8.

We deemed the number of annotations of relations to be too small to train a deep learning model efficiently. Instead, we adapted the corpus for linear-chain CRF, aligning with

Table 8. Distribution of most frequent relations in transformed corpus.

Relations	# relation	Ratio
ident (identity)	1,777	50.00
inclui / incluido (includes / included by)	717	20.17
ocorre_em / sede_de (location)	195	5.49
vinculo_inst (affiliation)	289	8.13
Total	3,554	100

Li et al. (36) and Collovini et al. (23), who extract relation descriptors for open relation extraction.

We adopted a semi-automatic approach that automatically identified relations between entities within the same sentence and manually annotated the relation descriptors. Focusing on descriptors with >150 explicit annotations, we excluded implicit relations of identity or inclusion, and prioritized affiliation ("*vinculo_inst*"), authorship ("*autor_de / obra_de*"), and placement ("*ocorre_em / sede_de*") among Group, Person, Place, and Title entities. Our annotation covered 71 affiliation relation descriptors, 51 authorship relation descriptors, and 35 placement relation descriptors.

Following the work of Li et al. (36), we replaced multi-token named entities with ARG1 and ARG2. In each entity pair, ARG1 represents a Title and ARG2 represents a Person in authorship relations, a Person and a Group in affiliation relations, and a Group and a Place in placement relations. In cases where sentences had multiple triples, we either split the sentence into multiple fragments to include all the relations or grouped the named entities as one.

To convert the datasets to a CSV (Comma-separated values) format, we used SpaCy to perform tokenization. We also provided linguistic information by performing part-of-speech tagging, and dependency parsing for each sentence. The output of the conversion is a file where each line represents a single word with the following fields: a word, a POS (Part-of-speech) tag, a dependency tag, and a relation classification label. Table 9 shows an example of the output of the parsing of the Arch.RE.Authorship dataset with the recognition of a relation of authorship between a work's title, "95 Teses" (95 Thesis), and a person, "Martinho Lutero".

Table 9. Example of parsing of Arch.RE.Authorship dataset.

Natural Language	As 95 Teses de Martinho Lutero
XML	As <EM ID="11" ENT="ARE2">95 Teses</EM> de <EM ID="12" ENT="E21" COREL="11" TIPOREL="autor_de">Martinho Lutero</EM>
Descriptors (XML)	As ARG1 <relation>de</relation> ARG2
Descriptors (CSV)	As DET det O ARG1 PROPN nsubj O de ADP case I-relation ARG2 PROPN nmod O

After integrating Collovini et al.'s (23) corpus, the final datasets contain 51 explicit authorship relations, 89 explicit affiliation relations, and 54 explicit placement relations.

Training of Machine Learning models

We applied a linear-chain CRF using the Python library *sklearn_crfsuite*[‖]. We considered two labels: I-relation and O, where a word labeled with I-relation is inside of a relation descriptor, and a word labeled with O is outside of a relation descriptor. We added context surrounding each token by including POS dependency tags for both the preceding and succeeding tokens. For the training, we used the default algorithm, gradient descent with Limited-memory BFGS (L-BFGS), with Elastic Net (L1 + L2) regularization of c1=0.1 and c2=0.01.

Evaluation of the models

We conducted a 5-fold cross-validation on open relation extraction with the contemporary dataset and performed a qualitative analysis with the archival dataset.

Contemporary general-domain dataset The 5-fold cross-validation of the Affiliation, Authorship, and Placement RE models resulted in an F1-score of 53.0% in the extraction of the affiliation relationship, 55.0% with the authorship relationship, and 66.2% with the placement relationship.

Twentieth-century archival dataset The CRF models were not able to correctly extract relations between named entities within sentences with the archival dataset.

There might be various reasons for this. First of all, the classifiers might be biased toward the corpora they were trained on as there is a small number of relation descriptors. Moreover, most of the training data is extracted from news articles. Naturally, news articles have a different linguistic requirement to letters and reports, which are the genre of the majority of the documents present in our sample. For example, structured reports and documents' covers tend to not have any syntactical relation between named entities. Most relationships in the dataset are either implicit or not expressed in the same sentence.

We give an example of identified relation descriptors with the affiliation relationship between an E21 and an E74 in Listing 1.

```
Ao Exmo. <relation>Presidente</relation> da <E74>
    Comissão</E74> d'Arrolamento dos Bens
do Palacio das Necessidades.
```

Listing 1: Example of an incorrectly extracted affiliation relationship from a letter in the dataset.

Listing 1 shows the phrase fragment "To Hon. Chairman of the Commission for the Inventory of Assets/ of the Palace of Necessidades.". The sentence contains an affiliation relationship between the Chairman and the Commission, with "of the" serving as the relation descriptor. However, the RE model incorrectly identified Chairman ("Presidente") as the relation descriptor, although it is an entity role.

Discussion

Flair models showed partial effectiveness for identifying common entities like people and places in 20th-century Portuguese archival digitized documents, but underperformed on other entity types due to data limitations, such as sparse annotations and noisy data. Notably, the smaller FlairEL embedding model matched or even outperformed FlairBBP for certain entity types in the Cultural Heritage domain.

Relation extraction was not fully effective for identifying relational facts in archival records' digital representations. CRF models failed to extract complete relational facts. In conclusion, while open relation extraction models were capable of recognizing relations, they did not reliably extract complete relational information.

The adaptation of general-domain datasets for CH tasks in 20th-century texts was marginally viable. Flair's noise handling ability did not translate effectively when trained on general-domain data, while RE models became ineffective due to domain mismatch, making it impossible to extract most relational facts.

[‖] https://sklearn-crfsuite.readthedocs.io/en/latest/ou

Conclusions and Future Work

The main goal of this work was to evaluate the effectiveness of general domain datasets for information extraction in the context of Portuguese Cultural Heritage, using an ontology-based information extraction system grounded in an archival linked data model.

BiLSTM-CRF models with Flair and Word2Vec embeddings demonstrated partial success in identifying entities in 20th-century Portuguese archival texts. However, relation extraction was insufficient to reliably extract relational facts.

We conclude that using general-domain datasets to fine-tune information extraction models for Cultural Heritage tasks showed limited viability. These findings highlight the need for domain-specific training to address historical archival data challenges.

For future work, we plan to expand the evaluation archival dataset to better understand the models' performance. Moreover, it would be interesting to explore alternative relation extraction solutions.

References

[1] Hyvönen E. *Publishing and using cultural heritage linked data on the semantic web*, volume 3. Morgan & Claypool Publishers, 2012. DOI: 10.2200/S00452ED1V01Y201210WBE003.

[2] Koch I, Ribeiro C and Lopes CT. ArchOnto, a CIDOC-CRM-Based Linked Data Model for the Portuguese Archives. In *Digital Libraries for Open Knowledge: 24th International Conference on Theory and Practice of Digital Libraries, TPDL 2020, Lyon, France, August 25–27, 2020, Proceedings*. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-54955-8, p. 133–146. DOI:10.1007/978-3-030-54956-5_10.

[3] CRM C. Definition of the CIDOC conceptual reference model. Technical report, ICOM/CIDOC CRM Special Interest Group, 2010. URL https://cidoc-crm.org/sites/default/files/cidoc_crm_version_5.0.2.pdf.

[4] Koch I, Lopes CT and Ribeiro C. Moving from ISAD(G) to a CIDOC CRM-based Linked Data Model in the Portuguese Archives. *J Comput Cult Herit* 2023; 16(4). DOI:10.1145/3605910.

[5] Ji H. *Information Extraction*. Boston, MA: Springer US. ISBN 978-0-387-39940-9, 2009. pp. 1476–1481. DOI: 10.1007/978-0-387-39940-9_204.

[6] Wimalasuriya C and Dou D. Ontology-based information extraction: An Introduction and a survey of current approaches. *J Information Science* 2010; 36: 306–323. DOI: 10.1177/0165551509360123.

[7] Vieira R, Olival F, Cameron H et al. Enriching the 1758 Portuguese Parish Memories (Alentejo) with Named Entities. *Journal of Open Humanities Data* 2021; 7. DOI:10.5334/johd.43.

[8] DigitArq. Memórias Paroquiais, 2021. URL https://web.archive.org/web/20211019034700/https://digitarq.arquivos.pt/details?id=4238720.

[9] Santos D, Seco N, Cardoso N et al. HAREM: An Advanced NER Evaluation Contest for Portuguese. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*. Genoa, Italy: European Language Resources Association (ELRA), pp. 1986–1991. URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/59_pdf.pdf.

[10] Nothman J, Ringland N, Radford W et al. Learning multilingual named entity recognition from Wikipedia. *Artificial Intelligence* 2013; 194: 151–175. DOI:10.1016/j.artint.2012.03.006.

[11] Zilio L, Finatto MJ and Vieira R. Named Entity Recognition Applied to Portuguese Texts from the XVIII Century. In Trojahn C, Finatto MJ, de Paiva V et al. (eds.) *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*, *CEUR Workshop Proceedings*, volume 3128. CEUR-WS.org, pp. 1–10. URL https://ceur-ws.org/Vol-3128/paper10.pdf.

[12] Semmedo JC. *Observaçoens Medicas Doutrinaes de Cem Casos gravissimos*. na officina de Antonio Pedrozo Galram, 1707.

[13] Lisboa JL, dos Reis Miranda TCP and Olival F (eds.) *Gazetas Manuscritas da Biblioteca Pública de Évora*, volume 1 (1729-1731). Évora: Publicações do Cidehu, 2002. DOI:10.4000/books.cidehus.3083.

[14] Paixão de Sousa MC. O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil. *Filologia e Linguística Portuguesa* 2014; 16(esp.): 53–93. DOI:10.11606/issn.2176-9419.v16ispep53-93.

[15] Cunha LFdC and Ramalho JC. NER in Archival Finding Aids: Extended. *Machine Learning and Knowledge Extraction* 2022; 4(1): 42–65. DOI:10.3390/make4010003.

[16] ADB. Arquivo Distrital de Braga . URL http://pesquisa.adb.uminho.pt/.

[17] DRABM. Arquivo e Biblioteca da Madeira . URL https://arquivo-abm.madeira.gov.pt/.

[18] Santos J, Cameron HF, Olival F et al. Named entity recognition specialised for portuguese 18th-century history research. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics, pp. 117–126. URL https://aclanthology.org/2024.propor-1.12/.

[19] Efremova J, García AM, Zhang J et al. Towards population reconstruction : extraction of family relationships from historical documents. In *First International Workshop on Population Informatics for Big Data*. pp. 1–9. URL https://dmm.anu.edu.au/popinfo2015/papers/2-efremova2015popinfo.pdf.

[20] Chantaraj P, Rungrattanaubol J and Na-udom A. Historical Relation Extraction from Buddhist Temple Documents of the Lanna Kingdom. *Journal of Computer Science* 2019; 15(9): 1320–1330. DOI:10.3844/jcssp.2019.1320.1330.

[21] Christou D and Tsoumakas G. Extracting Semantic Relationships in Greek Literary Texts. *Sustainability* 2021; 13(16). DOI:10.3390/su13169391.

[22] Santos D, Mamede N and Baptista J. Extraction of family relations between entities. In *INForum*. Citeseer, pp. 9–10.

[23] Collovini S, Machado G and Vieira R. A Sequence Model Approach to Relation Extraction in Portuguese. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 1908–1912. URL https://aclanthology.org/L16-1301.

[24] Freitas C, Mota C, Santos D et al. Second HAREM: Advancing the State of the Art of Named Entity Recognition in Portuguese. In Chair) NCC, Choukri K, Maegaard B et al. (eds.) *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Valletta, Malta: European Language Resources Association (ELRA). ISBN 2-9517408-6-7, pp. 3630–3637. URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/412_Paper.pdf.

[25] Collovini S, Neto JFS, Consoli BS et al. IberLEF 2019 Portuguese Named Entity Recognition and Relation Extraction Tasks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019) co-located with 35th Conference of the Spanish Society for Natural Language Processing (SEPLN 2019)*. pp. 390–410. URL https://ceur-ws.org/Vol-2421/NER_Portuguese_overview.pdf.

[26] Akbik A, Blythe D and Vollgraf R. Contextual String Embeddings for Sequence Labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*. Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 1638–1649. URL https://aclanthology.org/C18-1139.

[27] Bhadauria D, Sierra-Múnera A and Krestel R. The Effects of Data Quality on Named Entity Recognition. In *Proceedings of the Ninth Workshop on Noisy and User-generated Text (W-NUT 2024)*. pp. 79–88. URL https://aclanthology.org/2024.wnut-1.8/.

[28] Tjong Kim Sang EF and De Meulder F. 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*. pp. 142–147. URL https://aclanthology.org/W03-0419.

[29] Dias M and Lopes CT. Optimization of Image Processing Algorithms for Character Recognition in Cultural Typewritten Documents. *J Comput Cult Herit* 2023; 16(4). DOI:10.1145/3606705.

[30] Dias M and Lopes CT. Consensual ArchOnto representation of 13 Portuguese Historical Archival Records based on their Digital Representations. Data set, 2023. DOI:10.25747/EADP-M943.

[31] Akbik A, Bergmann T, Blythe D et al. FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 54–59. DOI:10.18653/v1/N19-4010.

[32] Santos J, Consoli B, dos Santos C et al. Assessing the Impact of Contextual Embeddings for Portuguese Named Entity Recognition. In *Proceedings of the 8th Brazilian Conference on Intelligent Systems*. pp. 437–442. DOI:10.1109/BRACIS.2019.00083.

[33] Lief E. *Deep Contextualized Word Embeddings from Character Language Models for Neural Sequence Labeling*. Master's thesis, Univerzita Karlova, Matematicko-fyzikální fakulta, 2019. URL http://hdl.handle.net/20.500.11956/105144.

[34] Segura-Bedmar I, Martínez P and Herrero-Zazo M. SemEval-2013 Task 9 : Extraction of Drug-Drug Interactions from Biomedical Texts (DDIExtraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*. Atlanta, Georgia, USA: Association for Computational Linguistics, pp. 341–350. URL https://aclanthology.org/S13-2056.

[35] Tjong Kim Sang EF and De Meulder F. 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*. pp. 142–147. URL https://aclanthology.org/W03-0419.

[36] Li Y, Jiang J, Chieu HL et al. Extracting Relation Descriptors with Conditional Random Fields. In *Fifth International Joint Conference on Natural Language Processing, IJCNLP 2011, Chiang Mai, Thailand, November 8-13, 2011*. The Association for Computer Linguistics, pp. 392–400. URL https://aclanthology.org/I11-1044/.