Named Entity Recognition (NER) in Historical Texts

Thomas Gaborieau
ENSAE
thomas.gaborieau@ensae.fr

Abstract

Named Entity Recognition (NER) in historical documents is a complex task due to OCR noise, non-standardized spelling, and scarce annotated resources. In this study, we evaluate multiple approaches for NER on French newspaper articles from the HIPE-2022 Le Temps dataset. We implement three categories of models: a logistic regression classifier using static FastText embeddings, two fine-tuned CamemBERT-based transformer models, and a zero-shot classifier using XLM-RoBERTa. Our results show that while static embeddings provide a fast baseline, they underperform on complex and contextual entities. Transformer-based models yield significantly better F1 scores, particularly for location and person entities. The Jean-Baptiste CamemBERT NER model slightly outperforms the base CamemBERT, confirming the benefit of prior task-specific pretraining. Zero-shot classification achieves high recall without fine-tuning, but precision is limited, especially for underrepresented classes such as organizations. We conclude that fine-tuned transformers are best suited for historical NER when resources allow, while zero-shot models offer a viable fallback for low-compute or preliminary filtering scenarios. Our analysis highlights the continued challenges of historical NER and the importance of model selection and data characteristics.

1 Introduction

Named Entity Recognition (NER) is a task in Natural Language Processing (NLP) that consists in identifying and categorizing named entities, typically persons, organizations, and locations inside unstructured texts. It plays a major role in a vast range of applications. While current NER systems achieves near-human performance on modern datasets in high-resource languages, their performance drops when applied to historical data. Historical documents present obstacles for the NER models. We can cite for example the presence of Optical Character Recognition (OCR) noise, outdated or region-specific vocabulary, non-standardized spelling, and others. As Ehrmann et al. (2022)[Ehr+22] write, "the simultaneous combination and magnitude" of these difficulties distinguish historical NER from other noisy-text domains such as user-generated content. To address these challenges, the HIPE (Identifying Historical People, Places and other Entities) initiative was proposed. The second edition of this shared task, HIPE-2022, was organized as part of the CLEF 2022 evaluation campaign. It focused on multilingual historical documents spanning the 18th to 20th centuries, looking at newspapers and classical commentaries, and involved both Named Entity Recognition and Linking (NERL). A main goal of HIPE-2022 was to measure the transferability and robustness of NER models in a multilingual setting and across time periods and document types. In this report, we experiment with different methods to approach NER in historical French-language newspaper articles from the HIPE-2022 Le Temps dataset. We start by reproducing a baseline using fine-tuned CamemBERT, a transformer model pretrained on contemporary French data. Then, we evaluate a Jean-Baptiste CamemBERT model pretrained on NER-specific data and we compare both to a logistic regression classifier using static FastText embeddings. Finally, we measure the abilities of multilingual Large Language Models (LLMs) in a zero-shot setting using the XLM-RoBERTa model. Our goal is to understand the extent to which modern methods—such as traditional embedding-based approaches and modern LLMs—generalize to historical data when there are low-resource and noisy conditions. We give anevaluation based on precision, recall, F1-score, and practical error analysis, yielding insights in the strengths and limits of each NER method on historical data.

2 Brief State-of-the-Art

Named Entity Recognition (NER) on historical data presents unique obstacles compared to modern texts. Historical French corpus often exhibit non-standardized orthography, obsolete vocabulary, and OCR noise, complicating the task. The HIPE initiative has been pivotal in advancing NER research on historical data. In the HIPE-2022 edition, different models were tested on multilingual historical data, including French datasets such as the newspaper ”Le Temps” with articles from the 19th and 20th centuries. For these data the best models achieved F1-scores of 0.66. Even better, the L3i team - from La Rochelle University - proposed a model that attained a F1-score of 0.808 on the ”hipe2020” French dataset, putting into light the efficiency of transformer-based models fine-tuned.

3 Methodology

3.1 Data Description

This project utilizes the French portion of the HIPE-2022 shared task corpus, specifically the letemps dataset. The letemps dataset consists of NE-annotated articles from two Swiss French-language newspapers, spanning the 19th and 20th centuries. It contains:

- 516 documents
- 466,600 tokens
- 11,045 entity mentions
- Three main NE types: person, location, organization
- Annotation format: CoNLL-style IOB with both coarse and fine entity types

This dataset was selected for its high-quality annotation and its relevance to the historical French language domain. It follows the same annotation guidelines as the HIPE-2020 dataset, ensuring consistent NE typology and tagging practices. One of the main challenges of the letemps dataset is its OCR noise, due to the digitization process of old newspapers. Around 20% of the entity mentions in the test set are affected by OCR errors, introducing variability and realism in evaluation scenarios. Moreover, the mention overlap between training and test sets is relatively high ($\tilde{2}$5.7%), making it suitable for controlled generalization studies while still offering a challenge for entity recognition systems.

Descriptive Statistics : After preprocessing and grouping tokens into sentences using the EndOfSentence markers, the data splits are:

Split	**Sentences**	**Tokens**	**Person**	**Location**	**Organization**
Training	14,051	$\sim$226,000	$\sim$2,465	$\sim$3,345	$\sim$1,190
Test	4,203	$\sim$70,000	352	575	77

Table 1: Summary statistics of the French HIPE-2022 *letemps* dataset after preprocessing.

The class distribution is notably imbalanced, with many sentences containing no named entities and a strong dominance of the “O” (non-entity) label. This imbalance, together with noisy input and sparse training examples for organization entities, constitutes a key challenge for learning and evaluation.

3.2 Experimental Setup

The objective of this study is to evaluate the performance of multiple Named Entity Recognition (NER) approaches on historical French newspaper data, specifically from the HIPE-2022 Le Temps dataset. Due to the noisy and linguistically distinct nature of historical texts, combined with limited computational resources, our methodological choices aim to evaluate three categories of models :

3.2.1 Static Embeddings with a Linear Classifier

As a baseline, we implement a lightweight NER pipeline using pretrained static word embeddings from **FastText** and a logistic regression classifier. This approach is motivated by its minimal computational requirements—embedding vectors can be preloaded once, and classification is performed efficiently using scikit-learn. While static embeddings lack contextual awareness and struggle with polysemy, they provide a fast and interpretable starting point. We hypothesize that this method will perform reasonably on surface-level entities like locations and common names, but poorly on entities requiring disambiguation or contextual cues. Nonetheless, it serves as a practical baseline that is robust to limited GPU memory.

3.2.2 Transformer-based NER with Fine-Tuning

We then fine-tune two transformer models for token-level NER using the Hugging Face Transformers library:

CamemBERT-base : A general-purpose transformer pretrained on modern French corpora. It provides a powerful contextual encoder but is not specialized for entity recognition.

Jean-Baptiste/camembert-ner : A CamemBERT model further fine-tuned for NER tasks, trained on contemporary labeled data. This model is expected to perform better than CamemBERT-base, particularly on common entity types.

These models are well-suited to the morphological complexity and variable word order in French, and they support subword tokenization which helps handle OCR artifacts and spelling variations in historical texts. However, their training and inference require significantly more computational resources than static embeddings. To mitigate this, we reduce batch sizes and limit training to three epochs.

3.2.3 Zero-Shot Sentence-Level Classification

Finally, we explore a zero-shot approach using the joeddav/xlm-roberta-large-xnli model, which supports multilingual classification. Here, we pose NER as a sentence-level classification task, predicting whether a sentence contains a person, location, or organization. This method avoids training altogether and requires only inference. While it cannot localize entity spans or distinguish multiple mentions, it offers an efficient triage mechanism to identify relevant sentences. We use it to evaluate sentence-level entity presence and analyze how well a general-purpose multilingual LLM transfers to historical domain-specific content.

4 Experimental Results

4.1 Static Embeddings (FastText + Logistic Regression)

Label	Precision	Recall	F1-score	Support
B-loc	0.52	0.58	0.55	591
B-org	0.00	0.00	0.00	79
B-pers	0.52	0.40	0.46	347
I-loc	0.56	0.03	0.06	151
I-org	0.00	0.00	0.00	130
I-pers	0.26	0.07	0.11	428
O	0.98	0.99	0.98	46742
Accuracy			0.97	48469
Macro avg	0.35	0.26	0.27	48469

4.2 Transformer-based Models

Model	Entity	Precision	Recall	F1-score
CamemBERT	LOC	0.60	0.86	0.71
CamemBERT	ORG	0.18	0.03	0.04
CamemBERT	PERS	0.60	0.78	0.67
Jean-Baptiste	LOC	0.63	0.86	0.73
Jean-Baptiste	ORG	0.15	0.16	0.15
Jean-Baptiste	PERS	0.56	0.74	0.64

4.3 Zero-Shot Classification

Zero-shot classification was applied to sentence-level entity recognition. Results from 200 sentences show effective identification of entity presence:

Example Sentence	Person	Location	Organization
Le ministère a en vain soutenu son système	0.01	0.77	**0.99**
Fabrique de J . CUEIU ' ILLOUD , à Rolle	0.57	**0.99**	0.99
Le grand-duc Constantin , ajoute - t - on...	**0.99**	0.97	0.99
Les conspirateurs condamnés à Pétersbourg...	0.81	**0.94**	0.99

5 Analysis of Results

The results of our experiments highlight important distinctions in performance between the different modeling strategies. These differences are largely influenced by the models' ability to handle the linguistic complexity and OCR-induced noise inherent in the historical French newspaper data, as well as their capacity to generalize across sparse and imbalanced entity distributions.

5.1 Static Embeddings with Logistic Regression

The FastText + Logistic Regression baseline performed surprisingly well for the non-entity (”O”) label, achieving nearly 99% accuracy. This is, however, reflective of dataset imbalance rather than classification ability. For named entities, performance was poor—especially for organization (ORG) and inside (I-) tags. This is expected: static embeddings lack context sensitivity and treat each token independently, which makes it difficult to resolve ambiguous words (e.g., "Paris" as a person vs. location). Moreover, the linear classifier cannot model sequential dependencies, making it particularly ineffective for multi-token entities (explaining the very low F1 for I-LOC, I-ORG, I-PERS). Despite these limitations, the method is computationally efficient and serves as a lightweight baseline when fine-tuning transformers is not possible.

5.2 Fine-tuned Transformer Models

Fine-tuning CamemBERT and Jean-Baptiste NER provided significant gains. Both models achieved F1 scores above 0.66 overall, a substantial improvement over static methods. These models benefit from subword tokenization (helpful for corrupted words), attention-based contextualization, and sequential modeling, all of which are critical for handling complex historical language.

Jean-Baptiste NER outperformed base CamemBERT slightly across all entity types, likely due to its prior training on NER-labeled French corpus. This supports the hypothesis that domain adaptation—even from modern domains—can still improve robustness when downstream data are scarce or noisy.

Interestingly, both models struggled with ORG entities. This can be attributed to:

- Their lower frequency in the dataset
- Greater lexical variability (e.g., press agencies, abbreviations)
- organization names often contain common nouns, making them harder to distinguish without prior knowledge

Additionally, performance was higher for LOC entities, likely because:

- Locations often appear in canonical forms (e.g., “Paris”, “Genève”) and follow common syntactic patterns (e.g., “à [location]”)
- They are more consistently annotated in the dataset

5.3 Zero-Shot Sentence-Level Classification

The zero-shot classifier (XLM-RoBERTa) offered a useful middle ground by identifying entity presence at the sentence level without requiring fine-tuning. Its recall for person entities (77.6%) and balanced performance on location (F1 0.68) suggest that LLMs pre-trained on multilingual NLI tasks retain strong semantic representations. However, precision for person was poor (29.5%), reflecting overprediction in sentences where semantic cues are weak. The complete failure to detect organization entities (F1 = 0.00) stems from :

- Severe class imbalance in the test subset (only 2 ORG-positive sentences)
- A lack of fine-tuning on entity-centered tasks

Despite these issues, the zero-shot approach is compelling for triaging historical text collections, where identifying sentences likely to contain entities can significantly reduce annotation or processing overhead.

5.4 Entity-Type Differences and General Challenges

Across all models, performance correlates strongly with entity frequency and regularity :

- LOC performs best due to high frequency and syntactic regularity
- PERS performs reasonably well, though variability in name structure and foreign names introduces challenges
- ORG remains the most challenging, reflecting annotation sparsity, lexical ambiguity, and label complexity

Finally, all models must contend with OCR artifacts and inconsistent spellings, which cause out-of-vocabulary issues, irregular tokenization, and label misalignment—especially detrimental in token-level models.

6 Conclusion

In this study, we explored a range of approaches for Named Entity Recognition (NER) on historical French data using the HIPE-2022 letemps dataset. Our goal was to assess how well modern NLP methods—ranging from simple embedding-based models to transformer-based architectures and zero-shot classifiers—can generalize to the challenges of noisy, domain-specific, and imbalanced historical data. We demonstrated that while static FastText embeddings combined with logistic regression provide a lightweight and computationally efficient baseline, they fall short on contextual and multi-token entity recognition, particularly for underrepresented entity types like organizations. Fine-tuning CamemBERT and its NER-specialized variant yielded significantly better performance, especially for locations and persons. Jean-Baptiste's NER model, with prior domain adaptation, showed modest gains over base CamemBERT, highlighting the value of task-specific pretraining. Our zero-shot experiments using a multilingual LLM (XLM-RoBERTa) showed promising recall and F1 for certain entity types without any fine-tuning. However, its inability to detect organizations and its low precision for persons expose the limits of zero-shot generalization in noisy and low-resource domains. These results confirm that historical NER remains a highly challenging task. Performance is sensitive to both entity type and model architecture, and all models are hindered by OCR errors and annotation sparsity. Nevertheless, transformer-based models, especially those adapted to NER, are the most reliable approach when computational resources are limited. Zero-shot methods may serve as a valuable complement for quick filtering or in annotation pipelines. To go further, we may explore hybrid approaches combining transformer predictions with static-embedding heuristics, OCR-correction preprocessing, or the integration of character-level models and nested NER frameworks to better capture historical linguistic variation and structure.

references

References

[Ehr+22] Maud Ehrmann et al. "Extended Overview of HIPE-2022: Named Entity Recognition and Linking in Multilingual Historical Documents". In: *Conference and Labs of the Evaluation Forum*. 2022. URL: https://api.semanticscholar.org/CorpusID:251471990.