Named Entity Recognition in Historical Italian: The Case of Giacomo Leopardi's Zibaldone

Cristian Santini[1,*], Laura Melosi[1,*] and Emanuele Frontoni[2,*]

[1] *Department of Humanities, University of Macerata, Macerata, Italy*

[2] *Department of Political Sciences, Communication and International Relations, University of Macerata, Macerata, Italy*

Abstract

The increased digitization of world's textual heritage poses significant challenges for both computer science and literary studies. Overall, there is an urgent need of computational techniques able to adapt to the challenges of historical texts, such as orthographic and spelling variations, fragmentary structure and digitization errors. The rise of large language models (LLMs) has revolutionized natural language processing, suggesting promising applications for Named Entity Recognition (NER) on historical documents. In spite of this, no thorough evaluation has been proposed for Italian texts. This research tries to fill the gap by proposing a new challenging dataset for entity extraction based on a corpus of 19th century scholarly notes, i.e. Giacomo Leopardi's Zibaldone (1898), containing 2,899 references to people, locations and literary works. This dataset was used to carry out reproducible experiments with both domain-specific BERT-based models and state-of-the-art LLMs such as LLaMa3.1. Results show that instruction-tuned models encounter multiple difficulties handling historical humanistic texts, while fine-tuned NER models offer more robust performance even with challenging entity types such as bibliographic references.

Keywords

Digital humanities, Historical documents, NER, Large language models, LLaMa, GliNER, Giacomo Leopardi, Zibaldone

1. Introduction

The increased digitization of the world's literary heritage opens new horizons for research in the field of Digital Humanities (DH). This research area has to confront some significant challenges that AI will face in the years to come. First of all, bibliographic resources are cultural products and, as such, they are subject to multiple interpretations and are described by libraries and research communities using various standards. This creates a wide variety of digitization approaches, which manifests in the multiplicity of digital edition projects and the creation of increasingly specific vocabularies and ontologies to describe these resources and their content [1]. From a digital perspective, this heterogeneity influences the ability of systems based on Machine Learning (ML) or Deep Learning (DL) techniques to understand and operate with this type of data. Nevertheless, the recent success of DL approaches in many fields opens up new possibilities for the DH. In specific, the advent of Large Language Models (LLMs), such as GPT [2] and LLaMa [3] pushed researchers to analyze their performances on domain-specific texts such as historical multilingual documents [4, 5].

The integration of language models into the analysis of humanistic texts is emerging as a promising field of study for Information Extraction (IE) and Named Entity Recognition (NER). Aladağ [6] uses LLMs in her study on the computational analysis of Evliya Çelebi's travels to identify semantic and thematic patterns, overcoming the linguistic challenges of Ottoman Turkish. In the proposed approach, the authors use ChatGPT for NER of entities such as proper names, people, places, and other categories within the text to calculate the most frequent entities. Similarly, the project described by Spina [7] aims to demonstrate how the use of AI tools, such as Transkribus and ChatGPT, can revolutionize the

EKAW 2024: EKAW 2024 Workshops, Tutorials, Posters and Demos, 24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024), November 26-28, 2024, Amsterdam, The Netherlands.

[*] Corresponding author.

c.santini12@unimc.it (C. Santini); laura.melosi@unimc.it (L. Melosi); emanuele.frontoni@unimc.it (E. Frontoni)
0000-0001-7363-6737 (C. Santini); 0000-0002-8893-9244 (E. Frontoni)

digitization and accessibility of Italian historical archives. In their study, the authors use ChatGPT for zero-shot recognition of references to persons and locations inside the Princes of Biscari correspondence. Overall, these studies highlight both the benefits and challenges of applying LLMs to humanistic texts, opening new perspectives for DH regarding the specialization of these tools for historical and literary documents.

Building on the state of the art, this research aims to quantitatively analyze the results of various NER algorithms based on language models by creating a new dataset for NER on historical Italian from a publicly-available resource: DigitalZibaldone [8], a digital edition of Leopardi's Zibaldone (1898). Written in several years by the Italian poet Giacomo Leopardi (1798 - 1837), the *Zibaldone* is a collection of more than 4000 notes on several topics, including philosophy, science, linguistics, politics, history and philology. The digital edition of this work, made by scholars and technologists at Princeton University, is an interesting resource for the evaluation of entity extraction tools for two reasons: first, it contains thousands of manually annotated references to people, places and bibliographic works linked to VIAF[1] and Wikidata[2] by domain-experts; secondly, the research notes of Giacomo Leopardi contained in this work represent a pre-digital example of an encyclopedic hyper-text, including ca. 10,000 internal and external references to historical figures, authors, literary works and other instances of humanistic knowledge [9]. To summarise, the contributions of this study are the following:

- to propose a new dataset for Named Entity Recognition and Linking in order to benchmark language models on historical humanistic texts;
- to understand how efficient LLMs can be in recognizing references to external entities in 19th century Italian scholarly notes and the actual challenges posed by these documents;
- to propose a new domain-specific NER model trained on historical humanistic texts;
- to provide a set of recommendations for integrating LLMs into the extraction of information from documents contained in archives of Italian authors.

This work is structured as follows. Section 2 introduces the current state of the art with respect to LLMs and NER, with a focus on applications for historical texts. Section 3 presents the dataset generated for the experiments, describes the different NER methodologies tested, and discusses in detail the evaluation metrics and tools used in the experiments. Section 4 reports the values obtained in measuring the models' ability to recognize entities from the Zibaldone. Section 5 provides a detailed analysis of the results obtained and final recommendations on the use of the analyzed models in literary digitization projects. Additionally, this section outlines future extensions of this study.

2. Related Work

The rise of LLMs has revolutionized NLP, enabling significant progress in tasks such as Named Entity Recognition (NER). Despite the impressive performance of LLMs for IE on modern texts [10, 11], their application to historical texts remains underexplored. Historical documents present unique challenges, such as language evolution, orthographic variation, and digitization errors, which hinder the direct application of models designed for contemporary language [12]. Currently, a growing body of research has attempted to leverage LLMs for historical NER. González-Gallardo et al. [4] explored the capabilities of ChatGPT for entity recognition in historical documents, applying it to diverse datasets such as NewsEye [13], hipe-2020 [14] and AJMC [15]. The study revealed that while LLMs could handle modern texts with some success, their performance on historical documents was inconsistent, particularly when dealing with the linguistic complexity of older texts. Problems such as entity overlap, code-switching between languages, and digitization errors, especially from optical character recognition (OCR), significantly degraded the model's accuracy. The authors concluded that task-specific training and better domain adaptation are required for LLMs to perform well on historical NER tasks.

[1] https://viaf.org/

[2] https://www.wikidata.org/

Similarly, another study by González-Gallardo and colleagues [5] evaluated open-source LLMs such as Llama [3] and Mistral [16] on historical NER tasks across a series of historical datasets in several European languages including French, German and English. The findings underscored the difficulty of recognizing entities in noisy, digitized texts, with models frequently misclassifying or failing to recognize domain-specific entities such as literary works. Moreover, LLMs struggled with the structural inconsistencies common in historical documents, such as incomplete or fragmented texts. Despite advancements in prompt engineering and multi-turn interaction modes, the performance of these models lagged significantly behind fine-tuned NER models, emphasizing the need for more specialized tools tailored to historical document analysis.

The study of historical texts in non-Western languages has also yielded similar results. Tang et al. [17] introduced the CHisIEC corpus for NER and Relation Extraction (RE) in ancient Chinese texts. The authors highlighted the linguistic diversity of ancient Chinese, which spanned multiple dynasties and included substantial variation in language use across time periods. Despite the use of specialized pre-trained language models fine-tuned on ancient Chinese, LLMs struggled with recognizing entities and faced particular challenges with the recognition of historical official positions and book titles. This study further emphasized the need for domain-specific language models, particularly in historical contexts where entity recognition is hampered by the lack of standardization in entity annotations and temporal variations in language.

In a different context, Spina [7] explored the use of AI tools like Transkribus and ChatGPT for transcribing and extracting entities from historical correspondence in the Biscari Archive, in Sicily. This project faced several difficulties due to the handwritten nature of the documents and their digitization into machine-readable formats. The integration of AI models like ChatGPT for zero-shot NER enabled the extraction of persons and locations from the transcribed texts, however the authors did not provide a thorough evaluation of the results and only a qualitative evaluation was provided. Lastly, Aladağ [6] applied the same chatbot to analyze Ottoman historical texts, particularly Evliya Çelebi's travelogue. While ChatGPT exhibited potential in extracting thematic and semantic patterns, the model struggled with the linguistic complexities of Ottoman Turkish, which incorporates elements from Arabic and Persian. The study concluded that current LLMs are ill-equipped to handle the intricacies of historical languages, and the lack of tailored models for such contexts limits the effectiveness of NER and other NLP tasks.

In all these studies, a common set of challenges emerges: noisy, digitized text, linguistic variation across time, and entity ambiguity or overlap. These challenges are particularly acute in historical texts, which lack the consistency and structure found in contemporary language corpora. While LLMs have shown promise in recognizing named entities in historical documents, their performance lags behind task-specific models, especially in the presence of OCR errors and complex, multilingual contexts. These limitations highlight the importance of developing models that are not only trained on historical language corpora but also capable of handling the variability and fragmentation inherent in such documents.

Notably absent from the existing body of research is a comprehensive study of NER in historical Italian texts, due to the lack of open and reusable datasets extracted from non-contemporary documents. Italian historical documents, ranging from medieval manuscripts to Renaissance [18] and post-Renaissance texts, exhibit significant orthographic and syntactic variations. Despite the need for accurate NER systems to support the DH in Italy, no large-scale, domain-specific corpus of texts currently exists for testing the performance of LLMs on historical Italian. Therefore, there is an urgent need to develop benchmarks and datasets for quantitative studies on the efficiency of LLMs in this domain. Such efforts would pave the way for more accurate entity recognition, fostering a deeper understanding of Italy's historical and cultural heritage.

3. Materials and Methods

3.1. Dataset

The objective of this research is to compare the performance of different open-source LLMs for Named Entity Recognition (NER) on historical Italian literary texts. To carry out this research, a new dataset for NER and Entity Linking (EL), was created based on a publicly available online resource: DigitalZibaldone[3]. This resource is structured as a website where each note of Leopardi can be accessed on a specific URI and is encoded using HTML. For example, by searching the page identified with URI https://digitalzibaldone.net/node/p2721_1 the user can visualize the first note in page 2721 of Leopardi's diary as an HTML page. In this edition, references to persons, locations and bibliographic works are encoded as links which redirect to the corresponding Wikidata entity, whenever present. If an entity is not present in Wikidata, an identifier will be provided using VIAF, the Virtual International Authority File.

Initially, 260 notes from the Zibaldone (pp. 2700-3000) written by Leopardi in 1823, one of the writer's most productive years, were identified as the evaluation dataset. Once these notes were identified, the HTML source code for each item was downloaded from the respective URI on the DigitalZibaldone platform using a web scraping algorithm developed in Python. A subsequent program, aimed at parsing the HTML file, was used to extract from each file the references to people, places and works defined through the hyperlinks present in the text. This algorithm produced two final CSV files: the first containing the text of each paragraph identified by an ID and cleaned of HTML formatting elements; the second containing all the annotations present in the 260 notes with the document ID, the position in the text of the annotated characters, the type of entity annotated (PER, LOC or WORK), and the link to the respective Wikidata or VIAF element. Examples of the two files contained in the evaluation dataset are shown in Table 1 and Table 2.

In order to compare the NER performance of large language models in a zero-shot setting with that of a smaller model fine-tuned on domain-specific data, a training dataset was extracted from two separate sections of the Zibaldone (pp.1000-2001 and pp. 3001-4000) by sampling notes with a total length less than or equal to 350 tokens and which contained at least a reference annotated. The reason to filter out longer texts is to let the model focus on documents in which information is concentrated in small sequences, since the Zibaldone is often composed of short notes which are semantically independent. After this filtering strategy, we obtained a total of 688 notes which were processed following the same pipeline described above. Table 3 reports the number of references in both the training and the evaluation dataset divided by class. In total, the dataset contains 2,899 references to people, places and literary works in the Zibaldone annotated by domain experts. The complete dataset is available on Zenodo [19].

In spite of the fact that this dataset has been sampled from a single author and that the NER annotations are limited to 3 coarse-grained classes, it still represents a valuable resource to test NER models on historical Italian for many reasons. First, the dataset contains ca. 850 references to literary works, which are one of the most challenging entity types to recognize in historical texts. Secondly, the data presents many domain-specific challenges from a linguistic point of view: one of them is that notes on the Zibaldone often contain quotes or expressions in Latin, Greek and French. Moreover, references to entities may be given by Leopardi with abbreviations, such as “Cic.” for “Cicero” or “Rep.” for “De Republica”. This is a frequent feature in scholarly writings and makes this dataset qualitatively similar to the AJMC corpus, a NER benchmark for 19th century classical commentaries in English, French and German [15].

[3] https://digitalzibaldone.net/

Table 1
Sample of a csv containing notes from DigitalZibaldone.

doc_id	text
https://digitalzibaldone.net/node/p2721_1	Anche il Gelli confessava (ap. Perticari Degli Scritt. del Trecento l. 2. c. 13. p. 183.) che la lingua toscana non era stata applicata alle scienze. (24. Maggio 1823.).

Table 2
Sample of a csv containing entity annotations.

doc_id	surface	start_pos	end_pos	identifier	type
https://digitalzibaldone.net/node/p2721_1	Gelli	9	14	Q518160	PER
https://digitalzibaldone.net/node/p2721_1	Perticari	31	40	Q3769747	PER
https://digitalzibaldone.net/node/p2721_1	Degli Scritt. del Trecento	41	67	viaf34613848	WORK

Table 3
Number of annotations divided by class in the dataset.

Dataset	PER	LOC	WORK
Training	1,093	407	635
Testing	492	61	211

3.2. NER Algorithms

3.2.1. Instruction-tuned LLM

The official LLaMa3.1 release provided by Meta was used in the experiments. In order to run the LLM on the setup configuration available, we used the 8B parameters instruction-tuned model available on Huggingface Transformers[4]. Since this model is not per se a NER model, we exploited the *in-context learning* capabilities of the conversational agent in order to generate answers which could be used to extract entities from a text. This was done by using two prompts:

- **Generative prompt**: the goal is to write a new version of the text with people, locations, and literary works which are annotated within the text with the following pattern: *left context <type>surface form</type> right context*
- **Extractive prompt**: given a passage from the Zibaldone, generates an ordered list of entities to be extracted from the text following the pattern: *<type 1>surface form 1</type 1>, <type 2>surface form 2</type 2>*

The instructions for both prompts were fed in Italian to mantain coherence with the input text. For more details, notebooks which show the step by step procedure to generate the output with the above mentioned prompts are available on GitHub[5].

After generating a response with the two prompts for each passage in our evaluation dataset, a post-processing algorithm was used to convert the output of the LLM into a CSV format which allows for automatic evaluation. First, a regular expression was used to identify from the answer all the annotations provided within angle brackets in order to get a list of pairs *(type, surface form)*. Then, a filtering algorithm was used to remove all the pairs which contained additional entity types, such as "DATE" or "ORG". Finally, the list of annotations was mapped to the start and end position of the tokens in the original passage and each annotation is converted to a tuple *(Note ID, Surface Form, Start, End, Type)*.

[4] https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct
[5] https://github.com/sntcristian/zibaldoned

3.2.2. Domain-specific NER Model

The GliNER library[6] was used to compare LLaMa3.1 with a smaller domain-specific NER model fine-tuned on our dataset. GliNER [20] are a family of NER models based on BERT [21] which were proposed as more efficient alternatives to large autoregressive models such as LLaMa [3]. This architecture aims to encode both spans (i.e., tokens) and entity representations in a shared embeddings space and the learning objective is to maximize the dot product between the span and class vectors. One of the advantages of these models is to be able to perform NER in a zero-shot setting by detecting entities of unseen classes.

In order to train a domain-specific model, a GliNER model trained for general-purpose NER in Italian with 90M parameters was used as base variant[7]. Thus, this model was trained in a cross-domain setting by fine-tuning it on the training portion of the Zibaldone. In order to carry the fine-tuning step, the dataset was pre-processed in order to convert entity types into the equivalent Italian words, e.g., "persona", "luogo" and "opera". After the pre-processing step, the dataset was split into training and validation using a $9/1$ ratio. The training was carried for four epochs with a learning rate of 5×10^{-6} for the NER components, i.e. the feed forward neural network and the span representations, a learning rate of 1×10^{-5} for the Transformer backbone, batch size 4 and weight decay 0.01. The low learning rate for the NER components was chosen in order to avoid catastrophic forgetting due to continual learning [22]. The training was conducted on a Dell7920 machine equipped with an Nvidia RTX A6000 GPU.

Furthermore, to understand the advantages of training a domain-specific model with manually annotated data, we compared the fine-tuned GliNER model with its base variant in a zero-shot evaluation. Since GliNER models allow for generalized NER, which is the task of recognizing entity without a closed vocabulary of entity types, we tested the 90M parameter model for Italian in a zero-shot setting by tuning it in order to detect entities described by the class "person", "place" and "work".

3.3. Evaluation Setup

Named Entity Recognition (NER) performance was assessed using both exact and fuzzy matching criteria, which is a common practice in evaluating NER models [23, 4, 5]. Both approaches require that the predicted class matches the reference class, but they differ in how they handle span detection. The exact matching criterion considers an annotation correct only when the predicted tokens perfectly aligns with the gold standard, while the fuzzy matching criterion allows for partial overlaps between the predicted and gold standard tokens. Using both criteria offers additional insights:

1. it helps assess how well the algorithms return correct annotations even if the boundaries of the mention are not perfectly identified;
2. examining the difference between exact and fuzzy matching in a per-class analysis allows to understand which entity types are more affected by boundary detection errors.

For both criteria (exact and fuzzy), precision, recall, and F1-score are calculated, with results being micro and macro-averaged across all classes as well as computed separately for each class.

4. Results

The results of the experiments are summarized in Table 4 and Table 5 respectively for the micro-averaged precision, recall and F1-scores and for all the metrics computed separately for each class and macro-averaged. The performance of four different algorithms - LLaMa3.1 (generative and extractive), GliNER (zero-shot and fine-tuned) - was evaluated using both exact and fuzzy matching. Both tables show that the GliNER fine-tuned model significantly outperforms the others across all metrics, achieving satisfactory results with a micro-averaged F1-score of 68.98% (exact) and 75.64% (fuzzy) on the overall

[6] https://pypi.org/project/gliner/

[7] https://huggingface.co/DeepMount00/GLiNER_ITA_BASE

dataset. This is particularly evident in the "person" class, where the model reaches a precision of 89,75% (exact) and 92% (fuzzy).

	Exact			Fuzzy		
	Precision	Recall	F1	Precision	Recall	F1
LLaMa3.1-8B (generative)	22,48	48,42	30,71	24,73	53,27	33,78
LLaMa3.1-8B (extractive)	37,06	29,06	32,58	44,07	34,55	38,74
GliNER (zero-shot)	30,6	50,79	38,19	35,33	58,64	44,09
GliNER (fine-tuned)	**75,15**	**63,74**	**68,98**	**82,4**	**69,9**	**75,64**

Table 4
Micro-averaged results of precision, recall and F1-score of the four algorithms computed on the evaluation dataset. For each evaluation metric, bold and underlined represent best and second best performance respectively.

		Exact				Fuzzy			
		PER	LOC	WORK	Avg.	PER	LOC	WORK	Avg.
Precision	LLaMa3.1-8B (generative)	28,97	9,00	12,69	16,89	29,85	9,00	18,07	18,97
	LLaMa3.1-8B (extractive)	56,87	15,85	15,19	29,30	61,02	17,07	28,92	35,67
	GliNER (zero-shot)	45,18	12,96	15,86	24,67	46,68	14,07	29,94	30,23
	GliNER (fine-tuned)	**89,75**	**81,25**	**44,50**	**71,83**	**92,00**	**81,25**	**63,50**	**78,92**
Recall	LLaMa3.1-8B (generative)	59,76	16,39	31,28	35,81	61,59	16,39	44,55	40,84
	LLaMa3.1-8B (extractive)	36,18	21,31	14,69	24,06	38,82	22,95	27,96	29,91
	GliNER (zero-shot)	60,98	57,38	25,11	47,82	63	62,29	47,39	57,56
	GliNER (fine-tuned)	**72,97**	**63,93**	**42,18**	**59,69**	**74,80**	**63,93**	**60,19**	**66,31**
F1	LLaMa3.1-8B (generative)	39,02	11,63	18,06	22,9	40,21	11,63	25,72	25,85
	LLaMa3.1-8B (extractive)	44,22	18,18	14,94	25,78	47,45	19,58	28,43	31,82
	GliNER (zero-shot)	51,9	21,15	19,45	30,83	53,63	22,96	36,7	37,76
	GliNER (fine-tuned)	**80,49**	**71,56**	**43,3**	**65,11**	**82,51**	**71,56**	**61,8**	**71,96**

Table 5
Precision, Recall and F1-scores computed per class and macro-averaged in both exact and fuzzy matching settings. For each class, bold and underlined represent best and second best performance respectively.

On the other hand, the LLaMa3.1 model with a generative prompt shows the weakest performance with an overall F1-score of 30.71% (exact) and 33.78% (fuzzy). This indicates that the generative capabilities of this model are not well-suited for NER tasks in historical texts, particularly when it comes to exact span detection. However, it is to notice that there is a slight difference in precision and recall when comparing the generative and extractive approaches, with the second achieving better precision. Indeed, in a per-class analysis, it is remarkable how LLaMa3.1 with an extractive prompt achieves the second best performance in terms of precision for both "person" and "location", outperforming the base variant of GliNER.

In terms of entity classes, the "person" class generally exhibits higher scores across models, with GliNER fine-tuned showing the best results with an F1-score of 82,51% and a precision of 92% (both fuzzy). However, the "work" class continues to present challenges, especially for non-trained models like LLaMa and GliNER zero-shot, where F1 drops below 20% in the exact matching setting. Additionally, the variance between exact and fuzzy matching is most pronounced in this class, indicating that models struggle to correctly identify boundaries for creative works, which often results in partial matches rather than exact ones.

5. Discussion and Conclusion

5.1. Discussion

The performance of the models shows a clear divide between those that are fine-tuned and those that rely solely on in-context learning or zero-shot capabilities. GliNER fine-tuned consistently outperforms the other models, demonstrating the value of domain-specific training when applied to complex literary texts. Its ability to identify entities with high precision and recall in both the "person" and "place" classes supports this observation. Particularly, for the "person" class, it reaches a precision of 92% in the fuzzy matching criterion, indicating robustness even in challenging text spans such as abbreviations. Conversely, the LLaMa3.1 model with a generative prompt emerges as the least performing model, with an F1-score of 30.71% (exact). This suggests that LLMs, even with their large parameter bases, may not be well-suited for tasks requiring precise entity recognition without fine-tuning, particularly in historical or literary contexts.

Among the entity classes, unexpectedly, the "place" class proves to be the most difficult to predict across LLMs. Instead models which are trained for general-domain NER, such as GliNER applied in a zero-shot setting, are achieving slightly better performance. This may be due to the fact that toponyms in Leopardi's notes may present lexical variations such as no-capitalized words, e.g. "italia", or abbreviations, e.g. "Venez." for Venice. Overall, the most challenging class, even for the best performing model, is the "work" class. This is a well-known problem in NER, due to the complex nature of references to creative works, which may often present abbreviations and lexical variations [24]. Specifically, a common error across all three models is the inability to distinguish bibliographic references expressed through the reference to the author of the work itself, which is a special ambiguity which is exemplified by Leopardi's frequent use of "Il Forcellini" to refer to the "Lexicon Totius Latinitatis" (1771). Such references require contextual information often absent in language models, leading them to classify these portions of text as references to people rather than literary works. Moreover, the "work" class exhibits the highest variance between exact and fuzzy matching, particularly in models like LLaMa3.1. This suggests that the boundaries for creative works are often misidentified, possibly due to the complexity of literary references and the inherent ambiguity of titles within the text. Fuzzy matching alleviates some of these issues, but precise identification remains a challenge.

5.2. Conclusion

In summary, this research proposed a comparative analysis of four NER algorithms based on language models for historical Italian literary texts. For this purpose, a new dataset was proposed that can be used by the entire research community to test the performance of various automated systems not only for NER but also for linking these references to Wikidata, i.e., Entity Linking (EL). This dataset, collected through the scraping of the online resource DigitalZibaldone, aims to become a new benchmark for measuring the effectiveness of automated systems supporting the creation of digital editions. Additionally, this study presents the application of four different NER approaches on this dataset and a quantitative analysis of the NER results obtained through standard metrics in information retrieval. The main takeaway from this research is that dealing with historical literary texts like Leopardi's Zibaldone presents unique challenges for NER tasks. One key difficulty is the recognition of domain-specific entities, particularly references to literary works, which are often referred to indirectly or with incomplete context. Models must also handle non-standard language use and orthographic variations that are typical of older texts, such as abbreviations and latin names.

In conclusion, what emerges from this research is that automatic entity recognition systems can be applied, albeit with several challenges, to humanistic texts. Fine-tuned models trained on domain data can be used to annotate literary texts, but human supervision is still necessary to verify the correctness and completeness of the annotations obtained. This suggests that NLP systems can be useful allies in digital editing projects, provided they are integrated into human-in-the-loop systems. On the other hand, large language models still seem far from being successfully applied to the analysis of humanistic texts, probably due to the predominance of web-derived texts in the training corpora of these models.

Future extensions of this work will focus on two main directions. The first will be to test models not only for entity recognition but also for EL, i.e., the disambiguation of a reference by linking it to its respective Wikidata element. This will aim to assess the feasibility of an architecture that automatically annotates references present in a literary corpus and semantically enriches them by linking them to an external *ad-hoc* knowledge base. The second direction, closely related to the first, is to propose a methodology for correcting and improving the performance of predictive systems for NER and EL by integrating them with semantic and logical conditions as well as statistical approaches. Since Knowledge Bases such as Wikidata [25] or Yago [26] contain multiple attributes and properties describing specific resources such as people, places and books, this information can be used to justify or contradict the predictions of an automatic entity extraction system by establishing temporal, spatial, or logical conditions to filter the elements of a knowledge base, as already proposed in other studies [23].

References

[1] P. Gaitanou, I. Andreou, M.-A. Sicilia, E. Garoufallou, Linked data for libraries: Creating a global knowledge space, a systematic literature review, Journal of Information Science 50 (2024) 204–244. URL: https://doi.org/10.1177/01655515221084645. doi:10.1177/01655515221084645, publisher: SAGE Publications Ltd.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, 2020. URL: http://arxiv.org/abs/2005.14165. doi:10.48550/arXiv.2005.14165, arXiv:2005.14165 [cs].

[3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. URL: https://arxiv.org/abs/2302.13971. arXiv:2302.13971.

[4] C.-E. González-Gallardo, E. Boros, N. Girdhar, A. Hamdi, J. G. Moreno, A. Doucet, Yes but.. Can ChatGPT Identify Entities in Historical Documents?, 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL) (2023) 184–189. URL: https://ieeexplore.ieee.org/document/10266291/. doi:10.1109/JCDL57899.2023.00034, conference Name: 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL) ISBN: 9798350399318 Place: Santa Fe, NM, USA Publisher: IEEE.

[5] C.-E. González-Gallardo, T. T. H. Hanh, A. Hamdi, A. Doucet, Leveraging Open Large Language Models for Historical Named Entity Recognition, 2024. URL: https://univ-rochelle.hal.science/hal-04662000.

[6] F. Aladağ, The Potential of GPT in Ottoman Studies: Computational Analysis of Evliya Celebi's Travelogue with NLP and Text Mining and Digital Edition with TEI, CULTURE 5 (2023).

[7] S. Spina, Biscari Epistolography. From Archive to the Website., DigItalia 18 (2023) 245–259. URL: https://digitalia.cultura.gov.it/article/view/3010. doi:10.36181/digitalia-00090, number: 2.

[8] S. Stoyanova, Working with the Digital Edition of Giacomo Leopardi's Zibaldone, magazén 4 (2023) 13.

[9] S. M. Stoyanova, Fragmentary narrative and the formation of pre-digital scholarly hypertextuality: G. Leopardi's Zibaldone and its hypertext rendition, in: Proceedings of the 3rd Narrative and Hypertext Workshop, NHT '13, Association for Computing Machinery, New York, NY, USA, 2013, pp. 1–6. URL: https://dl.acm.org/doi/10.1145/2462216.2462218. doi:10.1145/2462216.2462218.

[10] X. Wang, W. Zhou, C. Zu, H. Xia, T. Chen, Y. Zhang, R. Zheng, J. Ye, Q. Zhang, T. Gui, J. Kang, J. Yang, S. Li, C. Du, InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction, 2023. URL: http://arxiv.org/abs/2304.08085. doi:10.48550/arXiv.2304.08085, arXiv:2304.08085.

[11] O. Sainz, I. García-Ferrero, R. Agerri, O. L. d. Lacalle, G. Rigau, E. Agirre, GoLLIE: Annotation Guidelines improve Zero-Shot Information-Extraction, 2024. URL: http://arxiv.org/abs/2310.03668. doi:10.48550/arXiv.2310.03668, arXiv:2310.03668.

[12] M. Ehrmann, A. Hamdi, E. L. Pontes, M. Romanello, A. Doucet, Named entity recognition and classification on historical documents: A survey, arXiv preprint arXiv:2109.11406 (2021).

[13] A. Hamdi, E. Linhares Pontes, E. Boros, T. T. H. Nguyen, G. Hackl, J. G. Moreno, A. Doucet, A Multilingual Dataset for Named Entity Recognition, Entity Linking and Stance Detection in Historical Newspapers, in: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21, Association for Computing Machinery, New York, NY, USA, 2021, pp. 2328–2334. URL: https://dl.acm.org/doi/10.1145/3404835.3463255. doi:10.1145/3404835.3463255.

[14] M. Ehrmann, M. Romanello, A. Flückiger, S. Clematide, Extended Overview of CLEF HIPE 2020: Named Entity Processing on Historical Newspapers, CEUR Workshop Proceedings (2020). URL: http://ceur-ws.org/Vol-2696/paper_255.pdf. doi:10.5167/uzh-200192, conference Name: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum Meeting Name: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum Number: 2696 Place: Thessaloniki, Greece Publisher: CEUR-WS.

[15] M. Romanello, S. Najem-Meyer, A Named Entity-Annotated Corpus of 19th Century Classical Commentaries., Journal of Open Humanities Data 10 (2024).

[16] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7B, 2023. URL: http://arxiv.org/abs/2310.06825. doi:10.48550/arXiv.2310.06825, arXiv:2310.06825.

[17] X. Tang, Q. Su, J. Wang, Z. Deng, CHisIEC: An Information Extraction Corpus for Ancient Chinese History, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ELRA and ICCL, Torino, Italia, 2024, pp. 3192–3202. URL: https://aclanthology.org/2024.lrec-main.283.

[18] C. Santini, M. A. Tan, O. Bruns, T. Tietz, E. Posthumus, H. Sack, Knowledge extraction for art history: the case of vasari's the lives of the artists (1568), CEUR Workshop Proceedings 3234 (2022).

[19] C. Santini, ZibaldonED: Silver annotations for Entity Disambiguation from Digitalzibaldone, 2024. URL: https://zenodo.org/records/14103094. doi:10.5281/zenodo.14103094.

[20] U. Zaratiana, N. Tomeh, P. Holat, T. Charnois, GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 5364–5376. URL: https://aclanthology.org/2024.naacl-long.300. doi:10.18653/v1/2024.naacl-long.300.

[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[22] Z. Chen, B. Liu, Continual learning and catastrophic forgetting, in: Lifelong Machine Learning, Springer, 2018, pp. 55–75.

[23] C.-E. González-Gallardo, E. Boros, E. Giamphy, A. Hamdi, J. G. Moreno, A. Doucet, Injecting Temporal-Aware Knowledge in Historical Named Entity Recognition, in: Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I, Springer-Verlag, Berlin, Heidelberg, 2023, pp. 377–393. URL: https://doi.org/10.1007/978-3-031-28244-7_24. doi:10.1007/978-3-031-28244-7_24.

[24] N. Jain, R. Krestel, Who is mona l.? identifying mentions of artworks in historical archives, in: Digital Libraries for Open Knowledge: 23rd International Conference on Theory and Practice of Digital Libraries, TPDL 2019, Oslo, Norway, September 9-12, 2019, Proceedings 23, Springer, 2019,

pp. 115–122.

[25] D. Vrandečić, M. Krötzsch, Wikidata: a free collaborative knowledgebase, Communications of the ACM 57 (2014) 78–85.

[26] F. M. Suchanek, M. Alam, T. Bonald, L. Chen, P.-H. Paris, J. Soria, Yago 4.5: A large and clean knowledge base with a rich taxonomy, in: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, pp. 131–140.