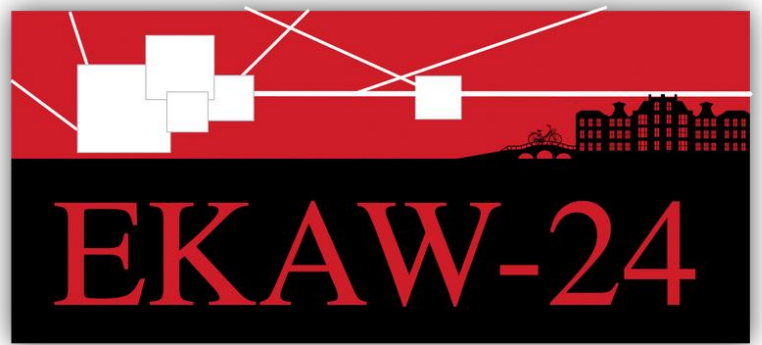
Named Entity Recognition in Historical Italian: The Case of Giacomo Leopardi's Zibaldone

Cristian Santini, Laura Melosi, Emanuele Frontoni

University of Macerata



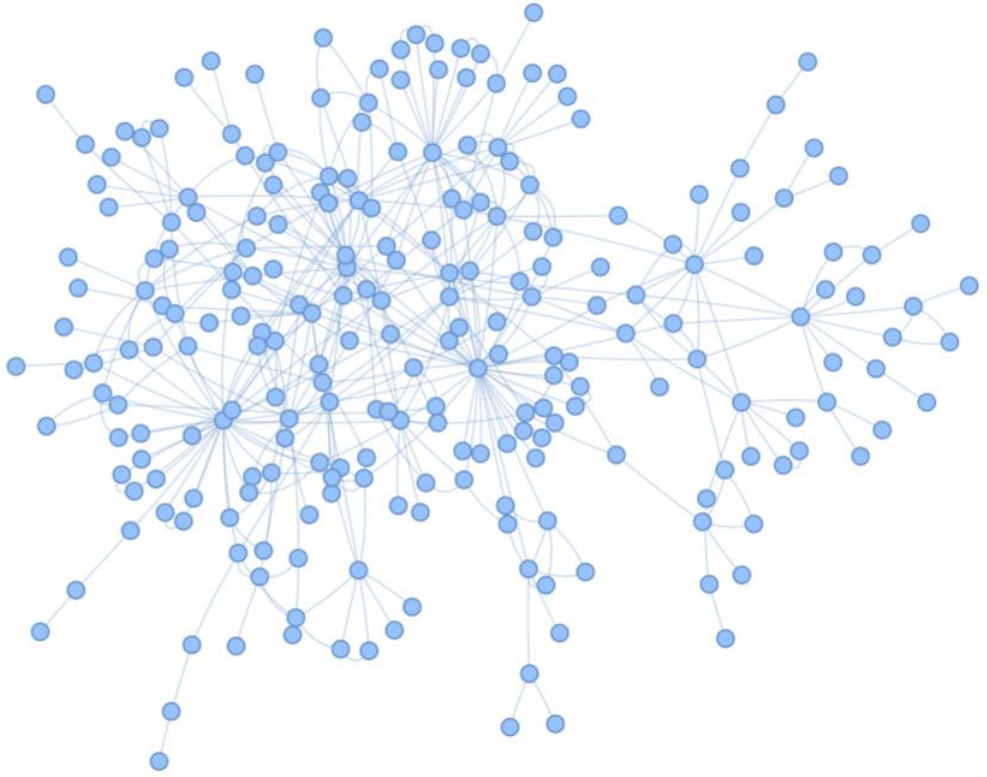
Context

- In 2017, the idea to catalogue and digitize the autograph manuscripts of Giacomo Leopardi was launched
- Since 2018, the University of Macerata has joined the project, renamed ‘Biblioteca Digitale Leopardiana’ [1]
- The goal is to launch a digital platform aimed at collecting images and metadata of the documents, as well as at enhancing Leopardi's texts and studies

[1] Melosi, L., & Marozzi, G. (2021). Il progetto Biblioteca Digitale Leopardiana: Per una catalogazione e digitalizzazione dei manoscritti autografi di Giacomo Leopardi. DigItalia, 16(1), Article 1. https://doi.org/10.36181/digitalia-00026

Leopardi's Digital Heritage

- Currently, more than 15,000 *facsimilia* of Giacomo Leopardi are digitized and publicly available on the web
- Efforts have been done to provide semi-diplomatic editions of his works [2] [3]
- Named Entity Recognition (NER) techniques are therefore needed to model facts expressed in natural language into a tailored Knowledge Graph



[2] Gioele Marozzi (Ed.), *Giacomo Leopardi.* Cambridge Digital Library. https://cudl.lib.cam.ac.uk/collections/leopardi/1
[3] Silvia Stoyanova and Ben Johnston (Ed.), *Giacomo Leopardi's Zibaldone di pensieri: a digital research platform*. https://digitalzibaldone.net/

Background: LLMs and Historical Texts

- Named Entity Recognition with ChatGPT on literary texts (ita,tr) [4] [5]
  - ✓ Applicable in absence of training data
  - X No quantitative evaluation
- Benchmarking LLMs on historical multilingual press articles [6]
  - X Vulnerable to noise and digitization errors
  - X Challenges of mitigating hallucinations and controlling output
  - X Worse performance than fine-tuned models

[4] S. Spina, Biscari Epistolography. From Archive to the Website., DigItalia 18 (2023) 245–259. https://digitalia.cultura.gov.it/article/view/3010.
[5] 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).
[6] González-Gallardo, C.-E., Hanh, T. T. H., Hamdi, A., & Doucet, A. (2024). Leveraging Open Large Language Models for Historical Named Entity Recognition. The 28th International Conference on Theory and Practice of Digital Libraries. https://univ-rochelle.hal.science/hal-04662000

Contribution

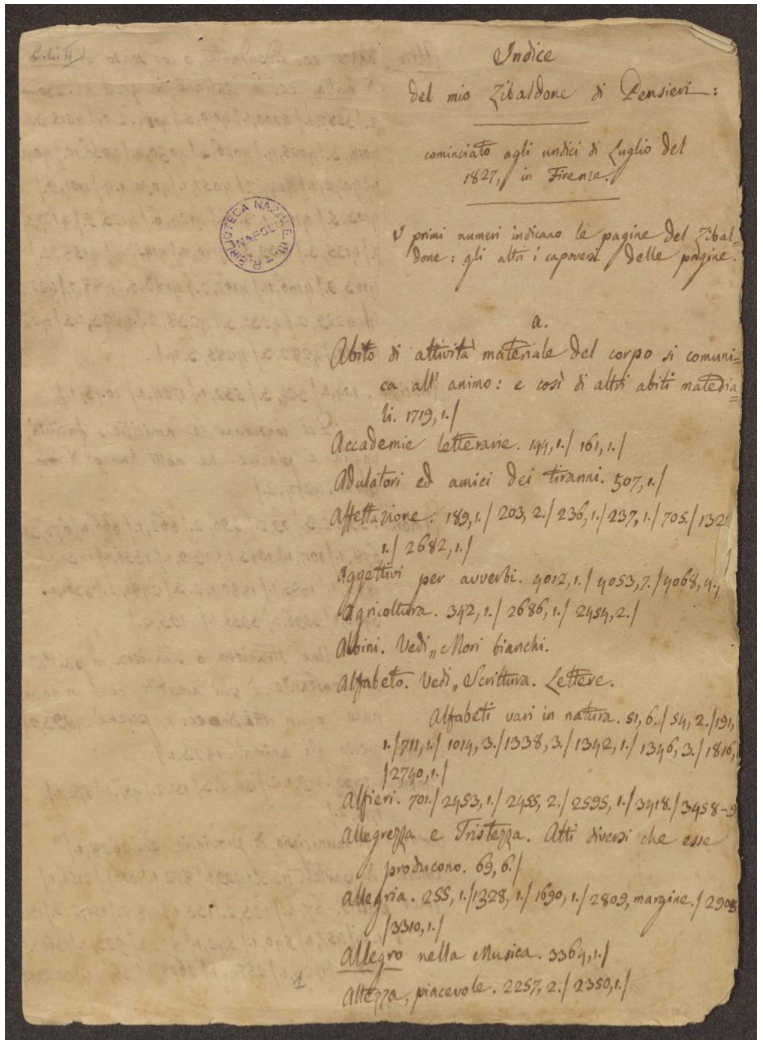
1. Propose a new dataset for Named Entity Recognition (NER) and Entity Linking (EL) extracted from Leopardi's *Zibaldone* (1817-1832)

2. Understand how efficient LLMs and fine-tuned models can be in recognizing references to named entities in Nineteenth century Italian humanistic texts

3. Give insights on the challenges of NER models and LLMs with Italian historical and humanistic texts

DigitalZibaldone

- DigitalZibaldone [3] is a digital edition of Leopardi's diary made by scholars in Princeton
- More than 4,000 encoded notes containing observations on a wide-range of topics (linguistics, history, philosophy, philology, politics, etc.)
- Over 10,000 references to people, places and literary works linked to Wikidata and VIAF and encoded in HTML


Indice
del mio Zibaldone di Pensieri:
cominciato agli undici di Luglio del
1827, in Firenze.
I primi numeri indicano le pagine del Zibal
done: gli altri i capoversi delle pagine.
a.
Abito di attività materiale del corpo si comuni=
ca all' animo: e così di altri abiti materia=
li. 1719, 1.
Accademie letterarie. 141, 1. 161, 1.
Adulatori ed amici dei tiranni. 307, 1.
affettazione. 189, 1. 203, 2. 236, 1. 237, 1. 705. 1321
1. 2682, 1.
Aggettivi per avverbi. 4012, 1. 4053, 1. 4064, 1.
Agricoltura. 342, 1. 2686, 1. 2454, 2.
Albini. Vedi „Mori bianchi.
Alfabeto. Vedi „Scrittura. Lettere.
Alfabeti vari in natura. 51, 6. 54, 2. 191,
1. 911, 1. 1014, 3. 1338, 3. 1342, 1. 1346, 3. 1816,
2740, 1.
Alfieri. 701. 2453, 1. 2455, 2. 2525, 1. 3118. 3458-
Allegrezza e Tristezza. Atti diversi che esse
producono. 63, 6.
Allegria. 255, 1. 1328, 1. 1690, 1. 2809, margine. 2905
3310, 1.
Allegro nella Musica. 3364, 1.
Allegro, piacevole. 2257, 2. 2350, 1.

[3] Silvia Stoyanova and Ben Johnston (Ed.), *Giacomo Leopardi's Zibaldone di pensieri: a digital research platform*. https://digitalzibaldone.net/

Markup of DigitalZibaldone

```
<p xmlns="http://www.w3.org/1999/xhtml" xmlns:tei="http://www.tei-c.org/ns/1.0">
<div class="node" id="p2721_1"><div class="nodemeta" id="p2721_1_meta"></div><b class="para_num">[2721,1]</b>NBSP Anche il <a
 class="person" href="https://digitalzibaldone.net/node/Q518160">Gelli</a> confessava (ap. <a class="person"
 href="https://digitalzibaldone.net/node/Q3769747">Perticari</a>
<a href="/node/viaf34613848">Degli Scritt. del Trecento</a> l. 2. c. 13. p. 183.) che la lingua toscana
non era stata applicata alle scienze. (24. Maggio 1823.).</div></p>
```

Data Sampling

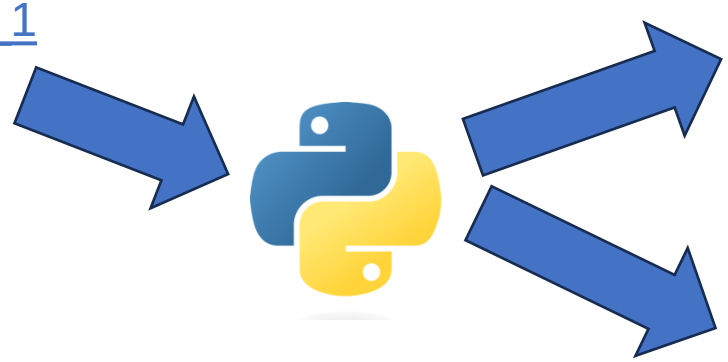
- **Evaluation dataset:** 260 notes sampled from pages 2700-3000
  - All written in 1823 → most productive year for Leopardi
- **Training dataset:** 688 notes sampled in page ranges 1000-2000 and 3001 – 4000 containing at least one entity
- Filtering documents in the training set longer than 350 tokens → keep texts with similar length

Data Pre-processing

« 2718,1 2721,2»

[2721,1] 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.).

https://digitalzibaldone.net/node/p2721_1



doc_id	text
p2721_1	Anche il Gelli confessava (ap. Perticari [...]

CSV

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

CSV

Dataset Statistics

Training

Statistic	Value
Number of documents	688
Number of annotations	2,135
Person Annotations	1,093
Location Annotations	407
Work Annotations	635

Evaluation

Statistic	Value
Number of documents	260
Number of annotations	764
Person Annotations	492
Location Annotations	61
Work Annotations	211

Models Tested

- **Instruction-tuned LLaMa3.1 8B [7]**: enhanced in-context learning capabilities for zero-shot NER
- **GliNER [8]**: BERT-based NER model trained on Italian texts to jointly learn entity and class representations
  - **Advantages:** can be evaluated both fine-tuned and in zero-shot setting

[7] Dubey, A., Jauhri, A., Pandey, A., Kadian, A., … Zhao, Z. (2024). The Llama 3 Herd of Models. arXiv. (arXiv:2407.21783). https://doi.org/10.48550/arXiv.2407.21783
[8] Zaratiana, U., Tomeh, N., Holat, P., & Charnois, T. (2023). GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer (arXiv:2311.08526). arXiv. https://doi.org/10.48550/arXiv.2311.08526

LLaMa Prompts

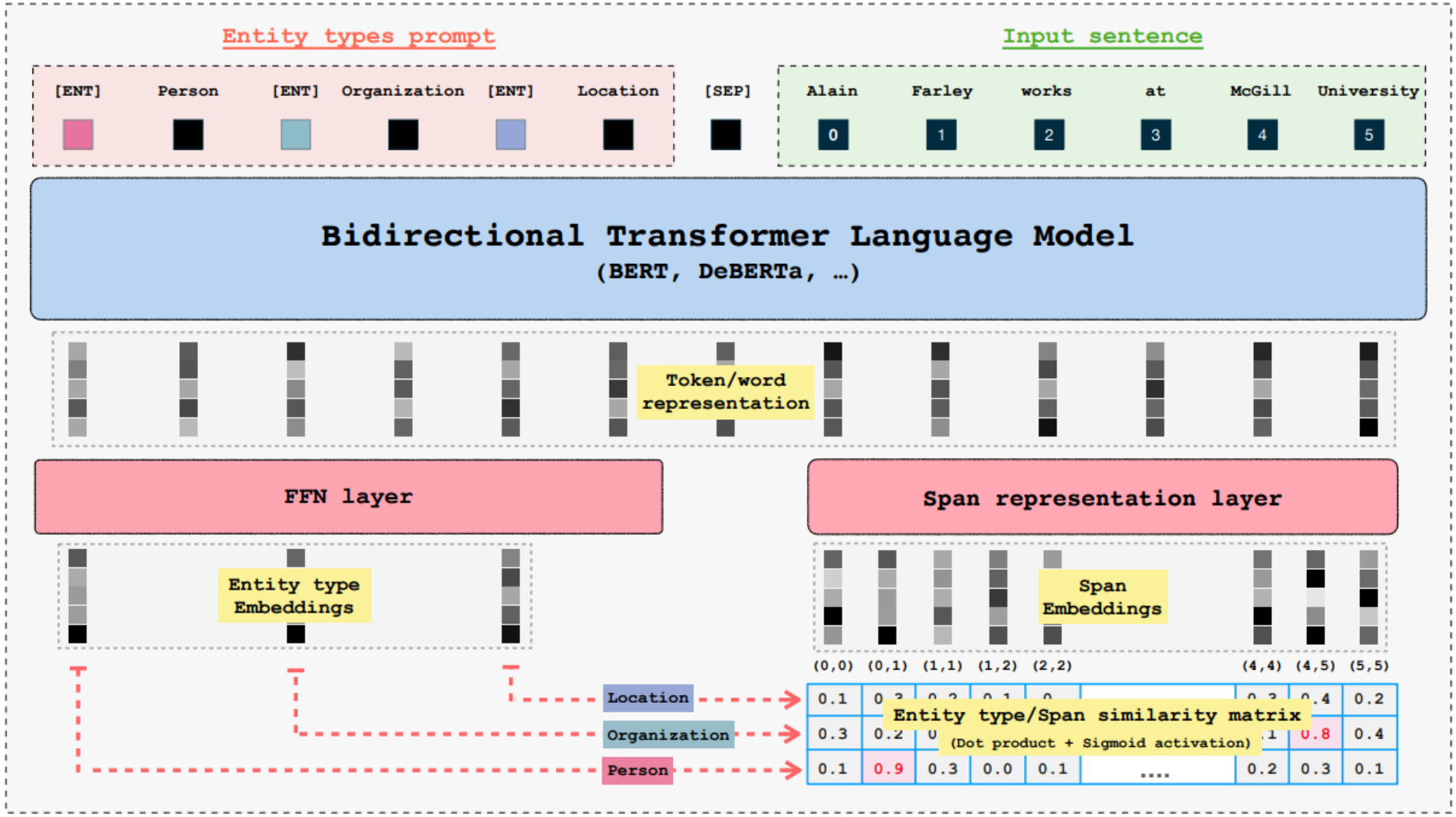
GENERATIVE PROMPT

Here's the annotated text: «The first example of <per>Dante</per>'s <work>Divine Comedy</work> is owned by <loc>Biblioteca Passerini-Landi </loc> in <loc>Piacenza</loc>.»

EXTRACTIVE PROMPT

Here is the list of entities: <per>Dante</per>, <work>Divine Comedy</work>, <loc>Biblioteca Passerini-Landi</loc>, <loc>Piacenza</loc>.

GliNER



[8] Zaratiana, U., Tomeh, N., Holat, P., & Charnois, T. (2023). GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer (arXiv:2311.08526). arXiv. https://doi.org/10.48550/arXiv.2311.08526

Evaluation Setting

Model	Setting 1	Setting 2
LLaMa3.1-instruct-8B	**Generative Prompt:** the LLM rewrites the annotated text based on a set of target classes	**Extractive Prompt:** the LLM extracts a list of entities from the text based on a set of target classes
GliNER_ita_base	**Zero-shot**: the model tries to identify unseen classes having been trained only on general texts	**Fine-tuned**: the model is pre-trained and fine-tuned on the training portion of the Zibaldone

Evaluation Metrics

- NER results computed using both *exact* and *fuzzy* matching
  - Allows to assess which class is more affected by boundary detection errors
- Precision, Recall and F1 computed for each algorithm
  - Micro-averaged across all classes
  - Computed separatedly for each class and macro-averaged

Micro-averaged Results

	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**

* Bold and underlined represent best and second-best results respectively

- All models used in a zero-shot setting are outperformed by the fine-tuned model
- LLaMa3.1 with a generative prompt shows the worst performance
- Using an extractive prompt increases significantly the precision for the LLM

Per-class Results

		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**

* Bold and underlined represent best and second-best results respectively

Key Insights

- Large pre-trained models like LLaMa require significant adaptation to handle the nuanced needs of NER on historical and literary texts
- All models are less prone to errors with respect to the person class
- Literary works remain the most challenging class due to varied references and complex naming conventions
- Challenges of lexical variations and abbreviations in surface forms (e.g., *Ven.* for Venezia)

Future Work

- **IE Benchmarking:** Collect further annotated documents to benchmark NER and EL on Italian texts from different periods
- **Entity Linking:** Extend the models' capabilities to not only recognize but also link entities to Wikidata elements
- **Model Improvement:** Develop methodologies for integrating semantic, logical, and statistical approaches to refine NER and EL predictions

Useful Links

Leopardi Knowledge Graph:

- https://sntcristian.github.io/leopardi_kg/

Zibaldone Entity Linking Dataset:

- https://zenodo.org/records/14103094

Cambridge semi-diplomatic edition:

- https://cudl.lib.cam.ac.uk/collections/leopardi/1

DigitalZibaldone:

- https://digitalzibaldone.net/

Biblioteca Italiana:

- http://www.bibliotecaitaliana.it/

Digital Library BNN:

- https://dl.bnnonline.it/handle/20.500.12113/4758

Leopardi Ecdosys:

- https://leopardi.ecdosys.org/it/Home/

WikiLeopardi:

- https://wikileopardi.altervista.org/wiki_leopardi/index.php?title=Wiki_Leopardi

Thank you!

Cristian Santini, Laura Melosi, Emanuele Frontoni

email: c.santini12@unimc.it