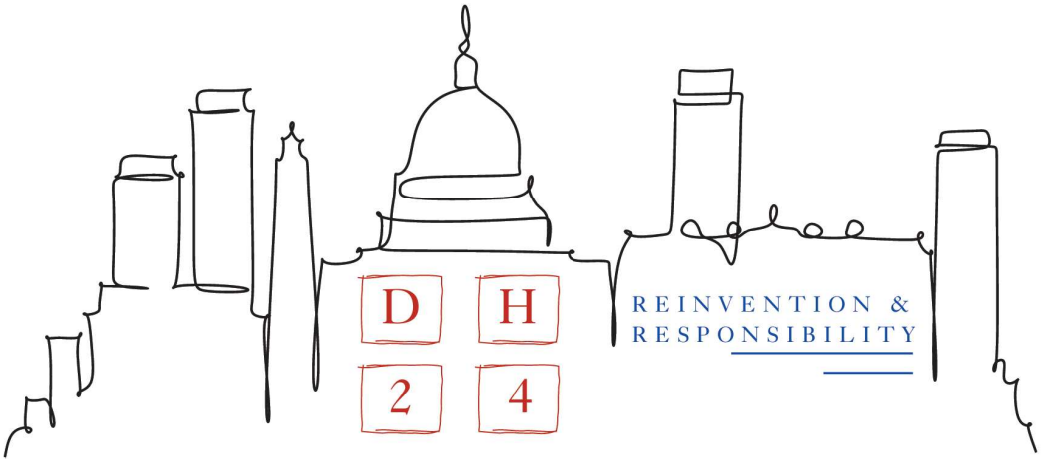


Book of Abstracts

August 6-9, 2024

Arlington, VA, U.S.A.

ALLIANCE OF DIGITAL HUMANITIES ORGANIZATIONS

Roy Rosenzweig Center for History and New Media

GEORGE MASON UNIVERSITY

Digital Humanities 2024: Book of Abstracts

Editors
Jajwalya Karajgikar, University of Pennsylvania Libraries
Andrew Janco, University of Pennsylvania Libraries
Jessica Otis, Roy Rosenzweig Center for History and New Media, George Mason University

Published by DH2024 Program Committee Chairs and Local Organizer Committee

Citation Recommendation

Karajgikar, Jajwalya/ Andrew Janco / Jessica Otis (eds.). Digital Humanities 2024: Book of Abstracts. VA 2024. DOI: 10.5281/zenodo.13761079

Leveraging Large Language Models to Generate a Knowledge Graph for Italian Literary Texts

XML

1. Introduction

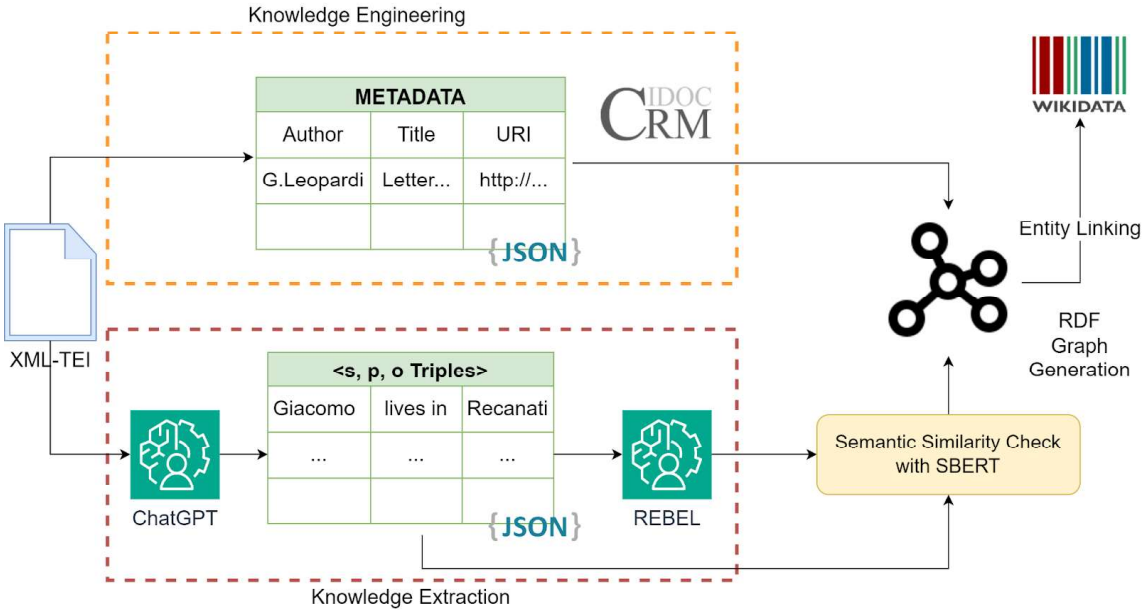
2. Recently, the advent of Large Language Models (LLMs) brought a paradigm shift in the realm of Natural Language Processing (NLP) and, as a consequence, in the Digital Humanities (DH). The fact that the massive amount of data from which these models are trained includes texts from several domains has shown to be an advantage for the capability of these models to accurately capture the meaning behind texts of several linguistic varieties and domains, including literary ones. As a consequence, NLP solutions based on LLMs have shown strong zero-shot performances on several tasks, such as Knowledge Extraction (KE) (Trajanoska et al., 2023). However, one of the main issues of LLMs is that they are prone to hallucinations. Since these models are trained through in-context learning, they are very capable of encoding relations between lexical entities but they are not able to instantiate properties between entities with declarative semantics, as in a Knowledge Graph (KG).

3. One of the crucial aspects in the digitalization of literary texts involves extracting entities and relations from a corpus. Entity Linking (EL) is aimed at spotting references to entities and concepts in a text and subsequently determining which entry in a Knowledge Base this reference should be linked to (Kolitsas et al., 2018). Relations, instead, are obtained via Relation Extraction (RE), an algorithm which decides whether two entities are connected by one of many relations specified using a controlled vocabulary or an ontology (Cabot & Navigli, 2021). However, besides (Brando et al., 2016), there is a lack of EL and RE approaches specifically suited for documents formatted in XML/TEI.

4. This research is aimed to provide a methodology for KE to be deployed on XML/TEI transcriptions of Italian literary texts, in order to extract formal, machine-readable representations of these documents which can be queried, explored and linked to external resources. The novelty of this approach is to address the limitations of current Relation Extraction models (RE) on Italian literary documents by employing ChatGPT (Openai, 2023) to transform unstructured texts into semi-structured formats, which can be easier interpreted by a general-purpose RE model. This abstract presents the following structure: the next section discusses the methodology, then a section will present the results obtained on a case study and the data considered as case study and future research steps will be outlined in the conclusion.

5. Methodology

6. In order to extract entities and relations from a given transcription, we propose to employ a pre-trained RE model, namely REBEL (Cabot & Navigli, 2021). This model takes as input a piece of text and returns a list of strings of the form <head, relation, tail> where head and tail are two entities identified in the text and the relation is a property from Wikidata which connects them. The advantage of this model is to extract only relations which are semantically defined in the Wikidata KG. We decided to employ REBEL which is a monolingual RE model trained on English instead of his multilingual variant mREBEL (Cabot et al., 2023) due to better performances during our experiments. However, REBEL does not work for Italian literary documents, therefore an intermediate step was needed to generate synthetic data in English based on our data. In order to circumvent this problem, we decided to adopt an instruction-tuned LLM, e.g., ChatGPT4, in a preliminary step to preprocess the transcriptions into a JSON file of the following format: [[entity1, relation, entity2], ...] where entities and relations are specified in English. For more details about the prompt used, source code is available on Github. Once a list of triples is produced as output by ChatGPT4, entities and relations are re-joined into a string and fed to REBEL in order to map the strings in English to RDF statements using Wikidata properties. In the last step, the output of REBEL is compared with the output of ChatGPT4 in order to find inconsistencies. This process is carried out by applying a threshold over the cosine similarity of the two strings encoded by a Sentence Transformer (Reimers & Gurevych, 2019). An overview of our approach is shown in Figure 1. In the final step, the extracted triples are integrated with possible metadata available in the input TEI transcription in order to disambiguate entities by leveraging already present identifiers from Wikidata, VIAF or GeoNames.



7.

7.1. Figure 1: A schema of our approach

8. Case Study and Results

9. In order to evaluate our approach we chose as case study the Leopardian manuscripts available in the Cambridge University Digital Library (CUDL). This dataset comprises 41 letters, fully transcribed in XML/TEI and related to Giacomo Leopardi. Since this is not a ground truth for evaluating KE methods, we compared the percentage of correct triples obtained with our approach with those obtained using a multilingual end-to-end RE model, i.e. mREBEL on the letters. While our approach generated 19 correct triples out of 38, with a 50% of accuracy, the considered baseline generated only 4 triples out of 40 were correct. This noticeable improvement in accuracy suggests the usefulness of integrating multiple components into a KE pipeline to adapt RE tools to multilingual documents.

10. Conclusions

11. To conclude, this research shows a novel approach to combine LLMs and fine-tuned models for RE on Italian literary texts, by adopting a LLM to preliminary pre-process texts into natural language triples in order to simplify the RE task for the fine-tuned model. After RDF statements are extracted, the quality of the output KG is to be evaluated. For this task, specific metrics for KG quality control (Wang et al., 2021) can be borrowed, in order to evaluate the semantic, syntactic and linking accuracy of the extracted triples and to compare different models or prompts.

12. References

13. **Brando, Carmen / Frontini, Francesca / Ganascia, Jean-Gabriel** (2016): "REDEN: Named Entity Linking in Digital Literary Editions Using Linked Data Sets", in: Complex Systems Informatics and Modeling Quarterly (7): 60. DOI: 10.7250/csimq.2016-7.04.

14. **Cabot, Pere-Lluís Huguet / Tedeschi, Simone / Ngomo, Axel-Cyrille Ngonga / Navigli, Roberto** (2023): "REDFM: a Filtered and Multilingual Relation Extraction Dataset", arXiv, <http://arxiv.org/abs/2306.09802> [31/05/2024].

15. **Huguet Cabot, Pere-Lluís / Navigli, Roberto** (2021): "REBEL: Relation Extraction By End-to-end Language generation", in: Moens, Marie-Francine / Huang, Xuanjing / Specia, Lucia / Yih, Scott Wen-tau (eds.): Findings of the Association for Computational Linguistics: EMNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics. 2370–2381, <https://aclanthology.org/2021.findings-emnlp.204> [31/05/2024].

16. **Kolitsas, Nikolaos / Ganea, Octavian-Eugen / Hofmann, Thomas** (2018): "End-to-End Neural Entity Linking", in: Proceedings of the 22nd Conference on Computational Natural Language Learning. Brussels, Belgium: Association for Computational Linguistics. 519–529, <https://aclanthology.org/K18-1050> [31/05/2024].

17. **Openai** (2023): "ChatGPT: Optimizing Language Models for Dialogue", in: archive.ph, <https://archive.ph/4snnY> [7/06/2024].

18. Reimers, Nils / Gurevych, Iryna (2019): “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”, arXiv, <http://arxiv.org/abs/1908.10084> [31/05/2024].

19. Trajanoska, Milena / Stojanov, Riste / Trajanov, Dimitar (2023): “Enhancing Knowledge Graph Construction Using Large Language Models”, arXiv, <http://arxiv.org/abs/2305.04676> [31/05/2024].

20. Wang, Xiangyu / Chen, Lyuzhou / Ban, Taiyu / Usman, Muhammad / Guan, Yifeng / Liu, Shikang / Wu, Tianhao / Chen, Huanhuan (2021): "Knowledge graph quality control: A survey", in: Fundamental Research 1 (5): 607–626. DOI: 10.1016/j.fmre.2021.09.003.

Cristian Santini (c.santini12@unimc.it), University of Macerata, Italy and Gioele Marozzi (g.marozzi4@unimc.it), University of Macerata, Italy and Emanuele Frontoni (emanuele.frontoni@unimc.it), University of Macerata, Italy and Laura Melosi (laura.melosi@unimc.it), University of Macerata, Italy

Author Index

Bart Holterman
Beatrice Vaienti
Bella White
Ben Miller
Benjamin Charles Nagy
Benjamin Gittel
Benjamin Krautter
Berat Kurar-Barakat
binxuan wang
Blair Tinker
Brecht Flora Marie Nijman
Brett Clark
Brian Croxall
Brittany Myburgh
Bruno Ministro
Byung-Chull Bae
Caitlin Rose Burge
Camilla Vang Østergaard
Cara Marta Messina
Carolina Villarroel
Cathy Moran Hajo
Cecil Krarup Andersen
Chang Liu
Changsong Wang
Chao Lang
Chao-Lin Liu
Charles Pletcher
Chiara De Bastiani
Chiara Palladino
Chris Hall
Chris Ruotolo
Christina Boyles
Christine Roughan
Christof Schöch
Christopher Kermorvant
Claudia Berger
Clelia R. LaMonica
Corinne Guimont
Craig Messner
Cristian Santini
Daniel Alves
Daniel Howlett
Daniel John Evans
Daniel Raffini
Daniel Riaño Rufilanchas
Daniele Metilli
Daria Vasyutinsky Shapira
David Neville
David Smith
David Valentine
Davide Pafumi
Devin Von Arx
Di Luo
Diana Milena Duarte
Diane Katherine Jakacki
Dian-Han Yang
Dinara Gagarina
Dominika Weronska
Dora Luise Muenster
Edison Weinstein
Eliane Schmid
Eliko Akashi
Elisa Eileen Beshero-Bondar
Elisa Ludwig
Elisabeth Guerard
Elise Heffernan
Elizabeth Losh

Ellen Charlesworth
Ellen Forget
Emiliano Andrés Calomarde
Emmanuel Château-Dutier
Emmanuelle Morlock
Enes Türkoglu
Enrico Agostini Marchese
Erik Fredner
Eunji Lee
Evelyn Lorraine Inman
Farnoosh Shamsian
Fatma Öncel
Fernanda Alvares Freire
Florian Cafiero
Florian Kessler
Florian Kräutli
Florian Windhager
Frank Fischer
Frédéric Clavert
Gabor Mihaly Toth
Gabriela Baeza Ventura
Gaby Barrios
Ganit Richter
Geoffrey Martin Rockwell
Geoffroy Noël
Georgia Panagiotidou
Georgii Korotkov
Gerald Kozicz
Gethin Rees
Gina Brandolino
Giulia Fabbris
Giulia Ferretti
Giuseppe Abrami
Glen Layne-Worthey
Goki Miyakita
Grace Cope
Graciana Ava Rudolfi
Grant Wythoff
Grażyna Urban-Godziek
Gregory Crane
Grzegorz Myrda
Guangwei Zhang
Guanwei Liu
Hale Sirin
Han Yan
Hanning Shao
Hao Zhu
Hassan El Hajj
Heidi Nobles
Hélène Huet
HeLing Sun
Henny Sluyter-Gäthje
Hisao Usui
Holly Chan
Hsin-Po Shen
Hugh Cayless
Hugo Scheithauer
Huiling FENG
Ian W. Scott
Isabel Galina Russell
Isabella Di Lenardo
Ismail Prada Ziegler
Iulianna van der Lek
Iuliia Iashchenko
Iwona Grabska-Gradzińska
J. Bern Jordan
J. Stephen Downie
Jacob Möhrke
Jakob Kusnick

James Adrian Balfour

James Cummings

Jami Ake

Jan Rybicki

Jana Smith Elford

Janis Dähne

Jason Dzubak

Jay Varner

Jeannette Schollaert

Jenna Stidwill

Jennifer Isasi

Jens Dörpinghaus

Jerald Lim

Jeremi K. Ochab

Jeremy Browne

Jesse P. Karlsberg

Jessica Monaco

Jianxuan Hong

Jiaqing LONG

Jimmy Medeiros

Jiying Kang

Joana Vieira Paulino

Joanna Katarzyna Hałaczkiewicz

Joaquin Rodriguez Cordeu

Joe Wicentowski

John Bradley

John McEwan

John O'Brien

Jonatan Jalle Steller

Jonathan Adam Clemons

Jonathan Girón Palau

Jörg Hörnschemeyer

José Calvo Tello

José Torres Álvarez

Joseph Anderson

Joseph Hiliary Nockels

Juan Cobo Betancourt

Judith Brottrager

Julia Flanders

Julia G. Polyck-O'Neill

Julia Jennifer Beine

Julian Häußler

Juliana Marques

June Wang

Justin Johnson

Justy Joseph

Kanji Kato

Katharina Hering

Katharine Emily Teykl

Katherine D. Harris

Katherine Ireland

Katherine Thurlow

Kathryn Ann Blizzard

Katie Lai

Katie R Muth

Ke Zhao

Keith Cheng

Keli Du

Keyanah Nurse

Kim Nayyer

Kiyonori Nagasaki

Konstantine Vlasis

Koraljka Kuzman Šlogar

Kristen Mapes

Kunal Vohra

L. Alberto Polo Romero

Lala Zuo
Laura Brannan Fretwell
Laura Rotunno
Lauren Coetzee
Lauren Klein
Laurie Allen
Lauryn Wilde
Lavanya Dahiya
Leanne Fan
Lee Bessette
Leticia Tobalina-Pulido
Lin Du
Linda Freyberg
LingYi Huang
Lisa Jean Baer-Tsarfati
Lisa Maren Poggel
Lisa Marie Rhody
Lisa Spiro
Lisa Tagliaferri
Lisandra Costiner
Lise Foket
Lorena Gauthereau
Lorna M Hughes
Luca Federico Cerra
Luca Giovannini
Lucas Avelar
Lucian Li
Lukas Rosenthaler
M Luisa Díez Platas
Maayan Zhitomirsky-Geffet
Maciej Eder
Maciej Kurzynski
Magdalena Turska
Mai Takahashi
Malte Vogl
Marc Alexander
Marc Lemke
Marcela Isuster
Marcello Vitali-Rosati
Marco Humbel
Marcus Bingenheimer
Marcus Pöckelmann
Margot Lise Mellet
Marguerite Jane Adams
Mari Kurokawa
Maria Beliaeva Solomon
Maria Fronczak
María Goicoechea
Maria Levchenko
María Ortiz Tello
Marie Anna Puren
Marie Flüh
Marija Blašković
Marine Clemence Avequin
Marisol Fila
Marjan Moosavi
Mark Andrew Algee-Hewitt
Mark Edwin Peterson
Markus Reisenleitner
Martha Attridge Bufton
Martin Grandjean
Martin Holmes
Martina Scholger
Mary Borgo Ton
Mary Naydan
Maryna Maslova
Masako Kubo
Masao Oi
Mason Jones

Mathias Zinnen
Matteo Romanello
Matthew Gold
Matthew Lavin
Megan Martinsen
Meghan Ferriter
Mehdy Sedaghat Payam
Meilong Chen
Melanie Conroy
Melanie Walsh
Melissa Jerome
Mengyuan Zhou
Meredith Graham
Meredith Martin
Merten Kröncke
Merve Tekgürler
Mia Ridge
Michael Sinatra
Michael Tiemann
Michaela Mahlberg
Michał Mrugalski
Michelle Meagher
Michikazu Kobayashi
Mila Aponte-González
Mila Oiva
MinHeng Lee
Mirerza González Vélez
Miriam Peña-Pimentel
Miriam Posner
Mmasibidi Setaka
Molly Taylor-Poleskey
Mona Elayyan
Morgan Pearce
Myeong Lee
Nanette Rißler-Pipka
Narayanamoorthy Nanditha
Natalia Ermolaev
Natalie Claire Seitz
Natalie Eloise Finlayson
Natalie Houston
Natalie Lorraine Liliensiek
Natalie McGartland
Natalie Nudell
Natalie Phillips
Nathan Kelber
Nathan Woods
Natsuko Nakagawa
Newman Chen
Nicholas Andrew Budak
Nicholas Lua
Nickoal Eichmann-Kalwara
Nicolas Quiroga
Nils Kellner
Nirmala Menon
Noelle Baker
Nuria Rodríguez-Ortega
Or Rappel Kroyzer
Orla Brid Delaney
Øyvind Eide
Pablo Ruiz Fabo
Paloma Vargas Montes
Pamella R. Lach
Panagiotis Papageorgiou
Paola Peratello
Paolo Verdini
Parham Aledavood
Pascal Belouin
Patrícia Esteves Reina
Patricia Ferreira-Lopes

Patricia García

Patrick Helling

Patrick J. Burns

Paul Pickering

Paweł Kamocki

Peizhen Wu

Peter Zhang

Petr Plechac

Pierre Willaime

Pilar Ramirez Restrepo

Pouyan Shahidi Marnani

Qilin Liu

Quinn Dombrowski

Quynh Tong

Rachel M Jacobson

Rachel Milio

Radomir Stanković

Raffaele Cioffi

Raffaele Viglianti

Raphael Schlattmann

Rebecca Nesvet

Rebecca Sutton Koeser

Rebekah Walker

Richard Khulusi

Richard Tzong-Han Tsai

Ricia Chansky

Riva Quiroga

Robert L. J. Shaw

Roberto Vargas

Robin Jegan

Robin Miller

Roch Delannay

Rongqian Ma

Rosalind J. Beiler

Ross Deans Kristensen-McLachlan

Ru Mo

Runze Song

Ryan Christopher Dubnicek

S.E. Hackney

Sam Blickhan

Samuel Ehrlich Backer

Sander Muenster

Sandra Toffolo

Sandy Aoun

Santiago Munoz Arbelaez

Sarah Bénière

Sarah Ellwein

Sarah Gilbert

Sarah Griebel

Sarah Hoover

Sarah Tew

Satoru Nakamura

Sayan Bhattacharyya

Sean Fraga

Sean Takats

Sebastian Barzaghi

Seiko Ochi

Servanne Monjour

Setsuko Yokoyama

Shani Evenstein Sigalov

Shanmugapriya T

Sharanya Ghosh

Shiyao Li

Shuang Xiao

Shumpei Katakura

Silke Schwandt

Simona Stoyanova
Smiljana Antonijevic
Song Chen
Sonia Marie Del Hierro
Sophia Booij
Sophia Martinez-Abbud
Spencer William Roberts
Stav Klein
Stefan Jänicke
Stefan Udelhofen
Stefanie Schneider
Stephanie Evert
Stephen Hart
Stephen Zweibel
Supreet Kaur
Susan Brown
Suzanne Alayne Moody
Suzanne Chase
Sven Najem-Meyer
Sydney Logsdon
Sylvia Arlene Fernandez
Talia Méndez
Tanvi Sharma
Tariq Yousef
Teresa Santa Maria
Terhi Nurmikko-Fuller
Thibault Clérice
Thora Marisa Hagen
Tobias Ebbrecht-Hartmann
Tobias Gradl
Todd Russell Hanneken
Tom Scheinfeldt
Toma Tasovac
Tomasz Umerle
Tong Li
Tong Wei
Tonisha Guin
Tonya-Marie Locke Howe
Torsten Hiltmann
Torsten Roeder
Toshinobu Ogiso
Trent Wintermeier
Tsz-Kin Chau
Tugce Karatas
Tyng-Ruey Chuang
Uhl Matthias
Ursula Lehmkuhl
Valeriia Korotkova
Vasundhra Dahiya
Vered Silber-Varod
Victoria Gerasimova
Victoria Van Hyning
Viktor J. Illmer
Vishal Hiteshbhai Patoliya
Vyshantha Simha
Walter Scholger
Wei Zhao
Wenqian Guo
Wenyi Shang
Whitney Sperrazza
William Reed Quinn
Winnie E. Pérez Martínez
Wouter Haverals
Wouter Paul Kreuze
Xiaochen Fang
Xiaotong Xu
Xingchen Zhang
XinXuan Wang
Xinyi Ding

Xuezhao Li

Ya-Chi Chan

Yael Levy

Yael Netzer

Yao Tong

Yohanna Joseph Waliya

Yoo Kyung Jeong

Yu Wang

Yuchen Yang

Yuerong Hu

Yufeng Han

Yumeng Hou

Yuqi Chen

Yuri Ishida

Yuta Hashimoto

Yuxiao Li

Zachary Lloyd

Zeynep Ecem Pulas

Zhenru Zhou

Zhuo Chen

Zoe LeBlanc

玉英 金