

Integrating Wikidata Entities into Narrative Graphs Using Large Language Models

Journal Title
XX(X):1-9
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/

SAGE

Emanuele Lenzi^{1,2} and Valentina Bartalesi¹

Abstract

Narratives are essential tools for articulating and sharing human experiences, particularly in scientific and cultural domains where they aid in the explanation of complex phenomena. In the context of a broader scientific effort in which Knowledge Representation and Semantic Web technologies are used to transform raw textual data into formal narratives, this paper explores the capability of Large Language Models (LLMs) to associate narrative entities (e.g. persons, locations, keywords) with the corresponding entity identifiers from Wikidata. We propose three LLM-based approaches for extracting and linking narrative entities and compare their performance against the JSI Wikifier, a state-of-the-art entity linking tool. The evaluation is based on a dataset from the H2020 MOVING project, which focuses on sustainability and value chains in European mountain regions. This study highlights the potential of LLMs to improve semantic annotation workflows and contribute to the automated generation of semantically enriched narratives.

Keywords

Large Language Models, Formal Narratives, Knowledge Representation, Semantic Web, Wikidata, Wikipedia

Introduction

A narrative serves as a means of meaningfully articulating life experiences (24) and represents a conceptual foundation for collective human understanding (27). Narratives, or stories, are utilised by humans to convey and communicate abstract experiences, express specific perspectives on a domain of interest, and share meanings across diverse communities (16; 6; 25). Narratives are particularly significant in the scientific context, where they are often used to explain complex phenomena and share specialised knowledge. To facilitate and standardise the creation of narratives in this context, we developed a semi-automatic workflow (1; 2) that leverages Semantic Web technologies (10) and the Knowledge Representation (KR) (14) to transform raw textual data into formal narratives. We define a narrative as a semantic network of events interconnected through meaningful relationships. Each event includes several entities, such as people, places, organisations, works, and concepts, which are uniquely identified by Internationalised Resource Identifiers (IRIs). The narrative is modelled using the Narrative Ontology (NOnt) (17; 3), a semantic framework developed by CNR-ISTI as an extension of the ISO standard CIDOC CRM (9) to formally represent narratives and their components. These narratives are published as OWL 2 DL (7) knowledge graphs, following the Linked Open Data paradigm (12), and visualised as story maps. Story maps are computer science realisations of narratives based on interactive online maps enriched with text, pictures, videos, data, and other multimedia information that can convey the emotional dimensions of a story and the places cited in it.

In this article, we evaluate the integration of Large Language Models (LLMs) into our workflow. Adopting a Semantic Web approach, we leverage nine LLMs to

automatically associate each entity mentioned in a narrative with an IRI, which uniquely identifies it, from Wikidata, which we adopt as an external reference knowledge base (26). To achieve this aim, we propose three different approaches that first extract entities (e.g., persons, locations, organizations, concepts) from raw narrative texts and then retrieve the Wikidata identifiers (i.e., QIDs) of the corresponding Wikidata entries, from which we are able to reconstruct the complete Wikidata IRIs. As a baseline for the comparison of the results of the three proposed approaches, we used the performance of the JSI wikifier (5). The JSI wikifier is a state-of-the-art framework that extracts key elements (i.e., named entities and keywords) from raw texts and links them to their corresponding Wikipedia pages.

As a real-world case study, we utilised English-language textual data collected as part of the H2020 MOUNTAIN Valorisation through INterconnectedness and Green growth (MOVING) project (18). These data are organised in an MS Excel file* and describe European mountain territories and value chains. Indeed, MOVING seeks to develop relevant policy to enhance the resilience and sustainability of mountain areas in response to climate change.

¹Institute of Information Science and Technologies (ISTI), National Research Council of Italy (CNR), Pisa, Italy

²Department of Information Engineering (DII), University of Pisa, Pisa, Italy

Corresponding author:

Valentina Bartalesi, Institute of Information Science and Technologies (ISTI), National Research Council of Italy (CNR), Pisa, Italy.

Email: valentina.bartalesi@isti.cnr.it

*https://github.com/AIMH-DHgroup/Linking_keywords_in_narratives_with_smaller_LLMs/blob/main/MOVING_VCs_DATASET_FINAL_V2.xlsx

The article is structured as follows: The Methodology section outlines the steps we followed, including the definition of three approaches to retrieve the Wikidata QIDs of narrative entities leveraging LLMs. The Task Definition section provides a detailed description of the task we want to perform and introduces the nine selected LLMs. Next, we present the dataset used as a case study and the gold standard corpus we created. The Results section presents the evaluation of the three approaches. In the Section Discussion, we report some considerations about the obtained results. Finally, the Conclusions section provides final remarks and outlines directions for future work.

Methodology

We adopted a bottom-up approach to identify and extract entities involved in narratives from raw texts, and to assign them Wikidata QIDs using LLMs. These QIDs enable the reconstruction of Wikidata IRIs, as QIDs are directly embedded within them. For example, the Carpathian Mountains are identified by the QID "Q1288", which is used to generate the corresponding Wikidata IRI, i.e., <https://www.wikidata.org/entity/Q1288>.

We began by providing the conceptual definitions of the terms *entity* and *narrative*, comparing our definitions with those found in the literature of Natural Language Processing (NLP), Narratology, and Psychology. Next, we selected the small open-source LLMs to use for our task. We also selected a dataset for the task and created a gold standard of manually annotated textual narratives, including the entities mentioned in the text and their corresponding Wikidata QIDs. To perform the task, we proposed three different approaches:

Approach 1: We instructed the selected LLMs to recognise the entities mentioned in the narrative texts and link them to their corresponding Wikidata entries by identifying their QIDs.

Approach 2: We instructed the selected LLMs to recognise the entities mentioned in the narrative texts. Then, we leveraged the Wikidata SPARQL endpoint (28) to fetch their corresponding Wikidata QIDs.

Approach 3: We instructed the selected LLMs to detect the entities mentioned in the narrative texts along with their corresponding Wikipedia titles. Then, we utilised the Wikipedia APIs[†] to obtain their Wikidata QIDs, using the identified Wikipedia titles as input.

For each approach, we calculate precision, recall, and F1 scores against the gold standard. These metrics were computed according to the following definitions:

- True Positive (TP): The model correctly identifies entities where the mentions in the text exactly match the annotated entities in the gold standard and links them to the correct Wikidata QIDs.
- False Positive (FP): The model predicts entities that either do not match the annotated entities in the gold standard or are linked to incorrect Wikidata QIDs.
- False Negative (FN): The model fails to identify entities annotated in the gold standard.

As a baseline for evaluating the performance of the three proposed approaches, we used the precision, recall, and F1

scores achieved by the JSI wikifier. The JSI wikifier is a state-of-the-art tool that identifies keywords in a textual document and links them with their corresponding Wikipedia entries.

Finally, we attempted to enhance the best results from the previous approaches by incorporating the Jaccard index (22). We used the Jaccard index to assess whether two entities identified by similar strings could be resolved as the same entity. Indeed, the Jaccard index is a metric used to measure the similarity and diversity of sample sets during the entity recognition step. Specifically, if two entities are represented by similar strings and exceed a certain similarity threshold, they may potentially correspond to the same entity.

Task Definition

The first step in defining our task is to clarify what we mean by *entity* and *narrative*.

We define entities as key textual elements extracted from narrative texts, often referred to as *keywords*. The meaning of the term *keyword* varies across different contexts and scientific disciplines. As highlighted by Notomo (19), a precise definition of a keyword has historically been lacking. The author cites Boyce et al. (4), who broadly define a keyword as “a surrogate that represents the topic or content of a document”. Notomo identifies four distinct roles that keywords can play: (1) terminology, referring to specialized lexical items within a particular domain; (2) topics, which include terms and labels within systematic concept systems like knowledge bases, for example, Wikidata; (3) index terms, highlighting significant concepts, events, or individuals, including named entities; and (4) summary terms, intended to provide a brief description of the content. In our study, we consider an entity to encompass all four roles outlined by Notomo. From this point forward, we will refer to these entities as *keywords*.

The meaning of *narrative* varies across different scientific fields. Narratives are often regarded as a tool for meaningfully organising life experiences (24) and as a conceptual basis for shared human understanding (27; 16). In psychology, narratives are considered fundamental to human life, helping individuals make sense of reality by arranging events into coherent stories (6; 25). In our study, we define a narrative from a KR perspective as a semantic network of spatiotemporal events connected by meaningful semantic relationships. By textual narrative, we mean a text that conveys a story consisting of a sequence of events constructed by a storyteller. Our goal is to perform a task that, leveraging the potential of LLMs, identifies keywords in textual narratives and retrieves the corresponding Wikidata QIDs, from which we can reconstruct the corresponding IRIs.

LLM Selection

Following an Open Science approach, we identified the two following key requirements that the LLMs must meet to perform our task:

[†]<https://en.wikipedia.org/w/api.php>

1. Open-source availability: The LLMs need to be open-source models, as our goal is to integrate them into our open-source-oriented workflow;
2. Hardware efficiency: The models should not require high-end hardware to ensure the experiment's repeatability, even with limited resources. Specifically, the experiment should be feasible on a desktop PC with an AMD Ryzen 7 5800X3D 8-core 16-thread processor, an Nvidia GeForce RTX 4090 24 GB GPU, and 32 GB of system RAM.

To identify the LLMs that can meet these two requirements, we used Ollama(20). Ollama is an open-source software offering some simple APIs and a library of pre-built LLMs, which can significantly reduce developers' time and effort. When conversing with a model, developers typically need to compose prompts in the specific syntax format required by that particular model. Ollama simplifies this process by automatically handling various prompt formats, allowing developers to test different models using a universal syntax.

To satisfy the requirements defined above, we decided to select nine small-sized, state-of-the-art open-source LLMs from Ollama, trained on different parameter scales, i.e., from 2 to 3.8 billion, from 7 to 9 billion, and from 13 to 14 billion. Whenever possible, we opted for their 8-bit quantised versions to achieve an optimal balance between precision and efficiency.

The selected small-sized models are the following:

1. Gemma 2 9b, one of the latest version of Google's open-source model with 9 billion parameters;
2. LLaMA 3 8b, one of the latest version of Meta's model with 8 billion parameters;
3. Mistral 7b, the latest version of the model of the Mistral AI team with 7 billion parameters that outperforms models trained with 13 billion parameters like LLaMA 2 13B.
4. Gemma 2 2b, one of the smallest version of the Gemma model, trained with 2 billion parameters;
5. LLaMA 3.2 3b, Meta's model with 3 billion parameters;
6. Phi 3.5, the latest version of Microsoft's model, trained with 3.8 billion parameters;
7. Deepseek 14b, a reasoning model that delivers performance comparable to OpenAI-01 across math, coding, and reasoning tasks;
8. Phi 4, the latest version of Microsoft's model, trained with 14 billion parameters;
9. LLaMa 2 13B, the latest version of a LLaMa model available in the 13-14 billion parameter range.

Dataset and Gold Standard Corpus

The dataset we created is a subset of the collection of the MOVING 454 narratives. These narratives were derived from the textual data in the English language collected in an unstructured MS Excel file, which was transformed using our semi-automatic workflow into 454 CSV files. Each CSV file represents a narrative about a specific mountain territory and contains 11 rows, each one corresponding to a narrative event(2). These textual events can be categorized

into three distinct groups. The first group focuses on descriptions of the territory's natural characteristics. In these events, keywords typically include the names of natural features relevant to the area, such as rivers or mountains. Examples include the Bregalnica River in North Macedonia or the White Mountains in Greece. The second group pertains to quantitative descriptions of the territory, covering aspects like geography, population, income, tourism, and employment. Keywords in this category often refer to measurable indicators, such as Gross Value Added or Per Capita Income. The third group highlights specific attributes of the region's products. Keywords in this category may refer to traditional products as well as the people or organisations involved in their production. Examples include the Pancretan Network of Organic Farmers of Aloe Vera, the Aloe Vera plant, and the Norwegian cheese Pustos.

To evaluate the results obtained by applying the three approaches described in Section Methodology, we created a dataset composed of a statistically significant subset of the 454 CSV files. To define this subset, we used the following sample size determination formula with finite population correction:

$$n_0 = \frac{Z^2 \cdot p \cdot (1 - p)}{MOE^2}$$

$$n = \frac{n_0 \cdot N}{N + n_0 - 1}$$

In this formula, n represents the target sample size adjusted for a finite population; n_0 is the initial sample size assuming an infinite population; Z is the Z-score that corresponds to the desired confidence level (we used Z-score =1.96 for the 95% confidence level); p represents the prior assessment (0.5 for uninformative conditions); MOE is the margin of error on the true error (5%), and N is the total number of events (population size).

Based on this methodology, 30 narratives - comprising a total of 330 events - were randomly selected, as this number was determined to be sufficient for a reliable performance evaluation. Finally, we created a gold standard corpus consisting of the 30 narratives that make up our dataset, which were manually annotated by a domain expert from the MOVING project. For each narrative event, the following information was annotated:

1. keywords
2. for each keyword, the corresponding Wikidata QID.

Note that if a keyword does not have any corresponding Wikidata entry, the value of its Wikidata QID is set to "null". The dataset and gold standard corpus are available on GitHub[‡].

Table 1 shows numerical data regarding the gold standard corpus, including the total number of textual events, the total number of annotated keywords in these events, the total number of unique annotated keywords, and the total number of annotated keywords without a corresponding entry in Wikidata.

[‡]https://github.com/AIMH-DHgroup/Linking_keywords_in_narratives_with_smaller_LLMs/tree/main/gold_standard

Table 1. Gold Standard data

Name	Events	Keywords	Unique keywords	Keywords with no Wikidata entry
MOVING	330	1960	464	306

Baseline Selection

To evaluate the three approaches described in the Methodology section, we first explored state-of-the-art frameworks capable of extracting textual fragments from documents and linking them to a knowledge base. Then, for each explored framework, we computed precision, recall, and F1 scores by comparing their predictions with the annotations in the gold standard. These metrics allowed us to identify the most effective framework to use as a baseline for evaluating our three approaches.

Since the MOVING gold standard contains annotations for both keywords (e.g., farmer) and named entities (e.g., North Macedonia), we selected only those frameworks capable of recognizing both element types. Consequently, we excluded entity-linking tools such as ReLik (21) or GENRE (8), which only identify and link named entities to a knowledge base. Furthermore, we also excluded tools like TENET (13), which perform joint entity and relation linking by connecting noun phrases to entities in general knowledge bases, as this approach falls outside the scope of our experiment.

The following three frameworks were selected for evaluation:

- The JSI wikifier, a framework that automatically identifies key text fragments in a document, including keywords and named entities, and links them to the corresponding Wikipedia entries. To ensure accuracy, it uses a PageRank-based algorithm that analyzes Wikipedia’s hyperlink structure, selecting the most relevant concepts based on the overall context of the document.
- TAGME (11), an automatic annotation framework that detects key terms in a text (referred to as spots, encompassing both keywords and named entities) and associates them with corresponding Wikipedia pages. TAGME uses an algorithm that evaluates the context and semantic coherence between the identified spots, selecting the most appropriate Wikipedia entry based on consistency with other spots in the text.
- Falcon 2.0 (23), an Entity Linking tool that connects both keywords and named entities in natural language texts to their corresponding elements on Wikidata. The process relies on linguistic techniques such as tokenization, Part-of-Speech tagging, and N-Gram tiling to identify keywords and named entities, generate candidates from Wikidata, and select the most relevant ones through a ranking system.

Using three Python scripts[§], we extracted and aggregated the textual descriptions for each event from the 30 selected narrative files in CSV format of the MOVING dataset. Then, the textual descriptions were processed using the three selected frameworks. Falcon 2.0 identifies keywords and directly assigns them the corresponding Wikidata QIDs.

Differently, the JSI wikifier and TAGME identify the keywords and assign them to the corresponding Wikipedia pages but are not able to directly retrieve the corresponding Wikidata QIDs. To retrieve the Wikidata QIDs of the keywords identified by the JSI wikifier and TAGME, we used the Wikipedia APIs.

Table 2 shows the precision, recall, and F1 scores of the JSI wikifier, TAGME and Falcon 2.0. The JSI wikifier demonstrates the highest precision among the selected frameworks (0.430), but it struggles to identify a sufficient number of keywords, resulting in a lower recall of 0.227. The JSI wikifier has also the best F1 score (0.297). TAGME, on the other hand, achieves a higher recall (0.590), finding more keywords annotated in the gold standard, but its low precision (0.141) suggests it generates many false positives. Furthermore, TAGME achieves the second best F1 score of 0.228. Falcon 2.0 performs poorly overall, with both precision (0.010) and recall (0.017) being extremely low, resulting in a very low F1 score (0.012).

Based on the F1 best score, we selected the JSI wikifier as the baseline.

Table 2. Precision, Recall and F1 score of the JSI wikifier, TAGME and Falco 2.0.

Model	Precision	Recall	F1
JSI wikifier	0.4304	0.2272	0.2974
TAGME	0.1414	0.5902	0.2282
Falco 2.0	0.0101	0.0175	0.0128

Results

This section presents the results (precision, recall, and F1 scores) obtained from each of the three approaches presented in Section Methodology.

First approach: Using LLMs to recognise the entities mentioned in the texts and retrieve the corresponding Wikidata QIDs

Using a Python script[¶], we extracted and aggregated the textual descriptions for each event from the 30 selected narrative files. Next, these 30 textual descriptions were passed to each LLM. The LLMs were tasked with identifying keywords and assigning to each of them the corresponding Wikidata QID. The LLMs’ outputs are formatted as JSON files, one for each CSV file. Each JSON file reports the keywords and QIDs identified by a given LLM. Thus, we obtained a total number of 270 JSON files. Finally, we calculated the precision, recall and F1 scores compared to the gold standard.

Table 3 presents the precision, recall, and F1 scores of the LLMs and the JSI wikifier, ranked by F1 score. The analysis reveals that the JSI wikifier ranks first, outperforming all LLMs. The LLMs performed poorly, with precision, recall,

[§]https://github.com/AIMH-DHgroup/Linking_keywords_in_narratives_with_smaller_LLMs/tree/main/Frameworks_for_baseline

[¶]https://github.com/AIMH-DHgroup/Linking_keywords_in_narratives_with_smaller_LLMs/blob/main/ollama.py

and F1 scores consistently below 0.1. This underperformance is primarily attributed to the LLMs’ tendency to generate fictional or incorrect Wikidata QIDs.

Table 3. Precision, Recall and F1 score of the JSI wikifier and the selected LLMs.

Model	Precision	Recall	F1
JSI wikifier	0.4304	0.2272	0.2974
Phi 4 14b	0.0117	0.0128	0.0122
LLaMA 3 8b	0.0089	0.0073	0.0080
Gemma 2 9b	0.0065	0.0066	0.0066
LLaMA 3.2 3b	0.0046	0.0052	0.0049
Gemma 2 2b	0.0007	0.0017	0.0010
Deepseek 14b	0.0006	0.0008	0.0007
LLaMA 2 13b	0.0000	0.0000	0.0000
Mistral 7b	0.0000	0.0000	0.0000
Phi 3.5	0.0000	0.0000	0.0000

Second approach: LLMs Integrated with Wikidata SPARQL Queries

To improve the results obtained in the first approach (Table 3), we modified the process by tasking the LLMs only with identifying the keywords in the texts, rather than directly linking them to Wikidata. The step of associating the detected keywords with their Wikidata QIDs was subsequently handled by a custom algorithm we developed[‡]. This algorithm performed SPARQL queries to the Wikidata SPARQL endpoint for each keyword identified by the LLMs, retrieving the corresponding Wikidata QIDs.

Table 4 shows the performances of the LLMs and the JSI wikifier, ranked by F1 score. Although LLMs’ performance improves slightly integrating the SPARQL-based Wikidata QID retrieval, it still lags behind the JSI wikifier. Phi 4 14b and Gemma 2 9b show the most significant improvement, with their F1 scores rising from 0.0122 to 0.1791 and from 0.0066 to 0.1597, respectively. Furthermore, Phi 4 14b achieves the highest recall among the LLMs at 0.1612, while Gemma 2 9b achieves the highest precision at 0.2263. Deepseek 14b ranks just below Gemma 2 9b with an F1 score of 0.1536. LLaMA 3.2 3b and Phi 3.5 still have F1 scores below 0.1, indicating a limited improvement in their entity linking capabilities. The other LLMs demonstrate F1 scores ranging between 0.11 and 0.14, showing a slight gain compared to the first approach.

Then, we aimed to improve the results in Table 4 by providing additional context to the LLMs. In particular, we supplied the selected LLMs with Natomo’s definition of “keyword”, which outlines the four roles described in Section Task Definition. By clarifying the concept of “keyword”, we sought to guide the LLMs in identifying them more accurately and meaningfully (15). Then, we applied the algorithm to retrieve the Wikidata QIDs for the keywords identified by the LLMs.

Table 5 shows the performance of the JSI wikifier and the LLMs, ranked by F1 score, when integrating keyword definitions. Precision, recall, and F1 show small improvements over the results in Table 4 for most models. Phi 4 14b and Gemma 2 9b exhibit a slight increase in precision (+0.016 and +0.035, respectively), recall

Table 4. Precision, Recall and F1 score of the selected LLMs and the JSI wikifier integrated with Wikidata SPARQL queries.

Model	Precision	Recall	F1
JSI wikifier	0.4304	0.2272	0.2974
Phi 4 14b	0.2016	0.1612	0.1791
Gemma 2 9b	0.2263	0.1233	0.1597
Deepseek 14b	0.1713	0.1392	0.1536
LLaMA 2 13b	0.1465	0.1393	0.1428
LLaMA 3 8b	0.1746	0.1064	0.1322
Gemma 2 2b	0.0973	0.1572	0.1202
Mistral 7b	0.1010	0.1309	0.1140
LLaMA 3.2 3b	0.1123	0.0898	0.0998
Phi 3.5	0.0310	0.0084	0.0133

(+0.093 and +0.058, respectively), and F1 score (+0.051 and +0.054, respectively). LLaMA 3 8b and DeepSeek 14b follow this trend, showing a small improvement in precision (+0.046 and +0.025, respectively), recall (+0.066 and +0.044, respectively), and F1 score (+0.061 and +0.036, respectively). LLaMA 3.2 3b and Mistral 7b show smaller improvements in precision (+0.017 and +0.010, respectively), recall (+0.040 and +0.003, respectively), and F1 score (+0.030 and +0.008, respectively). In contrast, Gemma 2 2b and LLaMA 2 13b show only slight improvements in precision (+0.002 and +0.004, respectively), along with a slight decrease in recall (-0.015 and -0.006) and F1 score (-0.003 and -0.001). Finally, Phi 3.5 shows a slight worsening in precision (-0.003) and a slight improvement in recall (+0.007) and F1 (+0.006). Overall, recall tends to improve more than precision for most models, except for Gemma 2 2b and LLaMA 2 13b. This suggests that providing keyword definition helps the models (except Gemma 2 2b and LLaMA 2 13b) correctly recognise more keywords.

Table 5. Precision, recall, and F1-score for the selected LLMs and the JSI wikifier, using provided keyword definition and the algorithm performing Wikidata SPARQL queries.

Model	Precision	Recall	F1
JSI wikifier	0.4304	0.2272	0.2974
Phi 4 14b	0.2182	0.2547	0.2351
Gemma 2 9b	0.2616	0.1817	0.2145
LLaMA 3 8b	0.2210	0.1725	0.1937
Deepseek 14b	0.1969	0.1836	0.1900
LLaMA 2 13b	0.1506	0.1331	0.1413
LLaMA 3.2 3b	0.1300	0.1305	0.1302
Mistral 7b	0.1114	0.1348	0.1220
Gemma 2 2b	0.0997	0.1418	0.1171
Phi 3.5	0.0274	0.0155	0.0198

Third approach: LLMs Integrated with Wikipedia APIs

Analysing the results obtained by combining the LLMs and the algorithm performing Wikidata SPARQL queries

[‡]https://github.com/AIMH-DHgroup/Linking_keywords_in_narratives_with_smaller_LLMs/blob/main/sparqlQuery.java

(Table 4 and Table 5), we observed that the algorithm frequently failed. This limitation arises because the SPARQL queries can only identify a Wikidata entry if the string that identifies the keyword exactly matches a Wikidata entry title or one of its alternative labels. For instance, the Wikidata entry “Slow Food”, an organisation for the protection of food biodiversity with QID “Q335823”, can only be retrieved if the string “Slow Food” is retrieved by the LLMs and then used in the SPARQL query. Alternative labels such as “Slow Food Foundation for Biodiversity” or “Slow Food Foundation”, although used in the textual events, and annotated in the gold standard, are not reported in Wikidata. As a result, these variations fail to yield a match.

Additionally, SPARQL queries lack flexibility in handling variations in word forms. For example, the string “farmers” (plural form) cannot be matched to the Wikidata entry “farmer” (singular form), as the query does not account for lemmatisation or morphological differences.

To address these challenges, we leveraged Wikipedia APIs, which offer flexible functionality to find a Wikipedia page corresponding to a textual keyword by using the page title as a parameter. Indeed, if a keyword extracted from a text does not exactly match a Wikipedia page title, the mechanism automatically attempts to resolve the match by redirecting to the correct page, interpreting the keyword string as a morphological variant or alternative form of the title. For instance, the keyword strings “Slow Food Foundation for Biodiversity” or “Slow Food Foundation” are successfully linked to the correct Wikipedia page titled “Slow Food”. Similarly, the entity string “farmers”, being a plural form, is accurately associated with the Wikipedia page “Farmer” in its singular form.

We refined our approach as follows: first, we supplied the selected LLMs with Natomo’s definition of “keyword”. Then, the LLMs were tasked with identifying the keywords in the CSV files. Next, for each identified keyword, the LLMs were prompted to determine the title of the corresponding Wikipedia page. Then, using the Wikipedia APIs and their redirect functionality, we retrieved the Wikipedia page associated with each title provided by the LLMs. Finally, starting from the retrieved Wikipedia pages, the APIs allow obtaining the corresponding Wikidata entries and their QIDs.

Table 6 presents the performance of the JSI wikifier and the LLMs, ranked by F1 score, using the revised approach. Comparing the results with Table 5, Gemma 2 9b and Deepseek 14b exhibit the highest improvements across all metrics, with increases in precision (+0.255 and +0.213, respectively), making them the models with the best and second-best precision among the LLMs. They also show gains in recall (+0.118 and +0.130, respectively) and F1 scores (+0.165 and +0.165, respectively), securing first and third place in the ranking, surpassing the JSI wikifier. Phi 4 14b follows closely, with improvements in precision (+0.146), recall (+0.103), and F1 scores (+0.126). It secures second place in ranking, overtaking the JSI wikifier as well. LLaMA 3 8b also benefits from the new approach, showing increases in precision (+0.165), recall (+0.086), and F1 scores (+0.116), placing it in fourth place, just above the JSI wikifier. Mistral 7b, LLaMA 2 13b, and LLaMA 3.2 3b exhibit similar trends, with improvements in precision (+0.130, +0.119, and +0.119, respectively), recall (+0.100,

+0.079, and +0.075, respectively), and F1 scores (+0.116, +0.096, and +0.095, respectively), positioning them very close to the JSI wikifier in the ranking. Gemma 2 2b and Phi 3.5 show more modest improvements, with precision increasing by +0.050 and +0.018, respectively, recall by +0.034 and +0.009, respectively, and F1 scores by +0.045 and +0.012, respectively. Overall, precision sees the highest improvements across most models, followed by F1 score and recall. This suggests that integrating Wikipedia APIs especially improves keyword-linking accuracy.

Table 6. Precision, recall, and F1-score of the JSI wikifier and the selected LLMs, with provided entity’s definition and integration of the Wikipedia APIs.

Model	Precision	Recall	F1
Gemma 2 9b	0.5169	0.2995	0.3793
Phi 4 14b	0.3645	0.3579	0.3612
Deepseek 14b	0.4099	0.3134	0.3552
LLaMA 3 8b	0.3857	0.2586	0.3096
JSI wikifier	0.4304	0.2272	0.2974
Mistral 7b	0.2411	0.2349	0.2380
LLaMA 2 13b	0.2700	0.2119	0.2374
LLaMA 3.2 3b	0.2487	0.2053	0.2249
Gemma 2 2b	0.1499	0.1755	0.1617
Phi 3.5	0.0460	0.0252	0.0326

Introducing the Jaccard Index to increase the LLM Results

As extensively described in the previous section, the third approach (i.e., LLMs Integrated with Wikipedia APIs) provides the best performance in terms of precision, recall, and F1 scores (see Table 6). Starting from these results, we decided to try further increasing them. To do this, we focused on the keyword recognition task performed by the LLMs. This task precedes the identification of the Wikidata QIDs associated with the keywords. This keyword recognition phase significantly impacts the accuracy of our task since, in the approaches presented in the previous section, if the string denoted by the keyword does not exactly match the keyword annotated in the gold standard, it is classified as a false positive. This affects the LLM precision values.

Upon analysing the keyword strings detected by the LLMs, we observed that some strings differ only slightly from those in the gold standard. The differences are due to two main causes. The first cause is the typos in writing the keyword names in the gold standard that are partially corrected by the LLMs, e.g. “Euopean Protected Geographical Indication” in the gold standard is retrieved and corrected by the LLMs as “European Protected Geographical Indication”. The second cause is the inclusion of additional words that identify a keyword, e.g. the string “Aloe Vera” is annotated as a keyword in the gold standard, instead, some LLMs identify the keyword string as “Aloe Vera plants”. To resolve this issue and allow for the identification of the correspondence between the two keyword strings, even when they do not exactly match character by character, we introduced the Jaccard index. The Jaccard index is a metric used to measure the similarity and diversity of sample sets.

Initially, we set the Jaccard index to 0.9 and then progressively lowered it to 0.7. Our observations revealed that by reducing the coefficient to 0.7, we obtained the best results in terms of precision, recall and F1 scores.

Table 7. Precision, recall, and F1 scores of the selected LLMs setting Jaccard index to 0.7.

Model	Precision	Recall	F1
Gemma 2 9b	0.5188	0.3050	0.3842
Phi 4 14b	0.3721	0.3799	0.3760
Deepseek 14b	0.4130	0.3237	0.3630
Mistral 7b	0.3160	0.3461	0.3304
LLaMA 3 8b	0.3878	0.2633	0.3137
JSI wikifier	0.4304	0.2272	0.2974
LLaMA 3.2 3b	0.2662	0.2279	0.2456
LLaMA 2 13b	0.2734	0.2172	0.2421
Gemma 2 2b	0.1564	0.1893	0.1713
Phi 3.5	0.0460	0.0252	0.0326

Table 7 presents the performances of the LLMs with the Jaccard index set to 0.7, demonstrating the impact of this more flexible approach. Comparing the results with Table 6, among the LLMs, Mistral 7b exhibits the largest improvement across all metrics, with precision increasing by +0.075, recall by +0.111, and F1 score by +0.092. It surpasses the JSI wikifier and LLaMA 3 8b in the ranking, placing fourth under Deepseek 14b. Phi 4 14b and Deepseek 14b exhibit a slight increase in precision (+0.0076 and +0.0031, respectively), recall (+0.0220 and +0.0103, respectively), and F1 score (+0.0148 and +0.0078, respectively). Gemma 2 9b and LLaMA 3 8b follow this trend, showing a small improvement in precision (+0.0019 and +0.0021, respectively), recall (+0.0055 and +0.0047, respectively), and F1 score (+0.0049 and +0.0041, respectively). Except for LLaMA 3 8b, which is overtaken by Mistral 7b in the ranking, all these LLMs maintain their positions. LLaMA 3.2 3b shows an improvement in precision (+0.0175), recall (+0.0226), and F1 score (+0.0207). These improvements allow it to surpass LLaMA 2 13b, which nonetheless experiences a small boost in precision (+0.0034), recall (+0.0053), and F1 score (+0.0047). Gemma 2 2b shows a slightly higher improvement in precision (+0.0065) and recall (+0.0138), leading to an F1 score gain of +0.0096, while remaining stable in the penultimate position in the ranking. Finally, Phi 3.5 does not improve or worsen.

Discussion

Our experiment, which leverages LLMs to extract keywords from textual narratives and link them to their corresponding Wikidata entries, demonstrates that while the LLMs initially underperformed compared to our baseline, the JSI wikifier, subsequent methodological refinements led to significant improvements. In particular, integrating LLMs with the Wikipedia APIs (third approach) proved to be a promising strategy for enhancing task performance. An analysis of the results in terms of precision and recall for this approach shows that the JSI wikifier outperforms most LLMs in precision, with the exception of Gemma 2 9b. Conversely, its recall is lower than that of the majority of the selected LLMs: Gemma 2 9b, Phi 4 14b, Deepseek 14b, LLaMA

3 8b, and Mistral 7b all achieve higher recall than the JSI wikifier. Regarding F1 scores, five of the nine selected LLMs outperform the baseline (see Table 7).

However, the overall low F1 scores achieved by both the LLMs and the JSI wikifier (with a maximum F1 score of 0.3842, reached by Gemma 2 9b) underscore the complexity of our task. Linking keywords to their corresponding Wikidata QIDs is particularly challenging when dealing with real-world datasets from specialised scientific domains. In our experiment, the narratives are derived from data written by experts in the bio-economic domain, and the gold standard keywords were annotated by one of these experts.

One of the most challenging aspects for both the LLMs and the JSI wikifier is the keyword identification step: correctly recognising which keywords have been annotated by an expert in the gold standard - before even determining their corresponding Wikidata QIDs - often proves to be a significant challenge. For instance, the keyword “tourism industry” (with the Wikidata QID Q9323634), annotated thirty times in the gold standard, is never identified by any of the LLMs. Conversely, the keyword “population density” (with the Wikidata QID Q22856), also annotated thirty times in the gold standard, is correctly identified but only by Phi 4 14b, the LLM with the highest recall. This discrepancy may be related to the textual context in which these keywords appear. For example, in the sentence “The population density (Inhabitants/km²) [Province] in the LAU is 82.56” some LLMs identify the numerical value (“82.56”) or the formula (“Inhabitants/km²”) as keywords rather than recognising the broader concept of “population density”. Additionally, some LLMs identify only the word “population” as a keyword instead of the full concept of “population density”. Furthermore, many events frequently contain domain-specific acronyms such as “LAU” (Local Administrative Unit) and “VC” (Value Chain), which are often not expanded within the textual descriptions. Despite this, they have been annotated in the gold standard by the MOVING expert. The LLMs can identify these acronyms; however, they often associate them with incorrect Wikipedia pages. For example, they link the acronym “VC” to the Wikipedia page for “Venture Capital” instead of “Value Chain”.

Finally, the expert annotated 306 keywords in the gold standard that refer to specific territorial products, events, companies, consortia, etc., but do not have a corresponding entry in Wikidata. These include, for example, the keyword “Cretan Diet Festival”, a festival in Crete dedicated to showcasing award-winning Cretan products. The LLMs often recognise these keywords and attempt to find a corresponding Wikipedia page title, even when no such page exists. Since Wikidata entries are linked to Wikipedia pages, the absence of a Wikipedia page also means there is no corresponding Wikidata entry. As a result, the number of false positives increases.

Conclusion

In this paper, we present an experiment that investigates the ability of nine Large Language Models (LLMs) to extract key entities from narratives and link them to their corresponding Wikidata entries, retrieving their QIDs. The

ultimate goal of this experiment is to retrieve the IRIs of these key elements to populate a narrative knowledge base. This experiment is part of a larger scientific effort to develop a semi-automatic workflow that combines Semantic Web technologies with Knowledge Representation to convert raw textual data into formal narratives. Following an Open Science-oriented approach, a key requirement for our experiment is that the selected LLMs be open-source and capable of running on hardware with limited resources, ensuring both reproducibility and accessibility. Based on this requirement, we selected the following LLMs: Gemma 2 9b, LLaMA 3 8b, Mistral 7b, Gemma 2 2b, LLaMA 3.2 3b, Phi 3.5, Deepseek 14b, Phi 4, and LLaMA 2 13B. As a baseline for our experiment, we use the results achieved by the JSI Wikifier, a state-of-the-art tool that links keywords to their corresponding Wikidata entries. We then compared the performance of the nine LLMs, along with the JSI Wikifier, against a manually annotated gold standard corpus. The dataset and gold standard corpus used in our study consist of a subset of narratives about European mountain territories, collected as part of the MOVING European project.

To conduct the experiment, we proposed three distinct approaches:

1. Prompt the LLMs to identify keywords mentioned in the texts and directly link them to Wikidata entries by retrieving their corresponding QIDs.
2. Prompt the LLMs to identify keywords in the texts, then use these keywords to query the Wikidata SPARQL endpoint to retrieve the corresponding QIDs.
3. Prompt the LLMs to identify keywords in the texts along with their corresponding Wikipedia titles. Then, use the Wikipedia API to retrieve the Wikidata QIDs using the Wikipedia titles as input.

While the results of the first approach showed that the LLMs underperformed compared to the JSI Wikifier, subsequent refinements to our methodology led to significant improvements. Specifically, integrating LLMs with the Wikipedia APIs (third approach) proved to be a promising solution for enhancing task performance. Analysing the ability of LLMs to identify keywords in the texts and their corresponding Wikipedia titles, the best performance was achieved by Gemma 2 9b, with a precision of 0.5188, a recall of 0.3050, and an F1-score of 0.3842. However, the F1-score obtained by Gemma 2 9b remains objectively low, underscoring the complexity of this task.

Future work includes evaluating our approach on the full dataset of 454 narratives collected within the MOVING project. Additionally, we plan to test our approach on narratives from various scientific domains to assess its interoperability and robustness. Another potential direction is fine-tuning the best-performing LLMs on domain-specific datasets, such as those focused on environmental studies or cultural heritage, to enhance their retrieval capabilities.

Acknowledgements

This work was partially supported by ITSERR (Italian Strengthening of the ESFRI RI RESILIENCE), funded under the MUR National Recovery and Resilience Plan funded by the European Union-NextGenerationEU. The funders had no involvement in the

study's design, data collection and analysis, publication decisions, or manuscript preparation.

References

- [1] Valentina Bartalesi, Gianpaolo Coro, Emanuele Lenzi, Pasquale Pagano, and Nicolo Pratelli. From unstructured texts to semantic story maps. *International Journal of Digital Earth*, 16(1):234–250, 2023.
- [2] Valentina Bartalesi, Gianpaolo Coro, Emanuele Lenzi, Nicolò Pratelli, Pasquale Pagano, Michele Moretti, and Gianluca Brunori. A semantic knowledge graph of european mountain value chains. *Scientific Data*, 2024.
- [3] Valentina Bartalesi, Emanuele Lenzi, and Nicolò Pratelli. A web tool to create and visualise semantic story maps. *Text2Story 2023 - Sixth Workshop on Narrative Extraction From Texts*. In *CEUR WORKSHOP PROCEEDINGS*, 2023.
- [4] Charles T. Meadow Bert R. Boyce and Donald H. Kraft. *Measurement in information science*. Academic Press, 1994.
- [5] Janez Brank, Gregor Leban, and Marko Grobelnik. Annotating documents with relevant wikipedia concepts. In *Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2017)*, Ljubljana, Slovenia, October 2017.
- [6] Jerome Bruner. The narrative construction of reality. *Critical inquiry*, 18(1):1–21, 1991.
- [7] World Wide Web Consortium et al. Owl 2 web ontology language document overview. 2012.
- [8] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.
- [9] Martin Doerr, Christian-Emil Ore, and Stephen Stead. The cidoc conceptual reference model: a new standard for knowledge sharing. In *Tutorials, Posters, Panels and Industrial Contributions at the 26th International Conference on Conceptual Modeling - Volume 83*, ER '07, page 51–56, AUS, 2007. Australian Computer Society, Inc.
- [10] John Domingue, Dieter Fensel, and James A Hendler. *Handbook of semantic web technologies*. Springer Science & Business Media, 2011.
- [11] Paolo Ferragina and Ugo Scaiella. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management, CIKM '10*, page 1625–1628, New York, NY, USA, 2010. Association for Computing Machinery.
- [12] Pascal Hitzler and Krzysztof Janowicz. Linked data, big data, and the 4th paradigm. *Semantic Web*, 4(3):233–235, 2013.
- [13] Xueling Lin, Lei Chen, and Chaorui Zhang. Tenet: Joint entity and relation linking with coherence relaxation. In *Proceedings of the 2021 International Conference on Management of Data, SIGMOD '21*, page 1142–1155, New York, NY, USA, 2021. Association for Computing Machinery.
- [14] Arthur B Markman. *Knowledge representation*. Psychology Press, 2013.
- [15] Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. Prompt engineering in large language models. In I. Jeena Jacob, Selwyn Piramuthu, and Przemyslaw Falkowski-Gilski, editors, *Data Intelligence*

- and *Cognitive Informatics*, pages 387–402, Singapore, 2024. Springer Nature Singapore.
- [16] Greg J McInerny, Min Chen, Robin Freeman, David Gavaghan, Miriah Meyer, Francis Rowland, David J Spiegelhalter, Moritz Stefaner, Geizi Tessarolo, and Joaquin Hortal. Information visualisation for science and policy: engaging users and avoiding bias. *Trends in ecology & evolution*, 29(3):148–157, 2014.
- [17] Carlo Meghini, Valentina Bartalesi, and Daniele Metilli. Representing narratives in digital libraries: The narrative ontology. *Semantic Web*, 12(2):241–264, 2021.
- [18] MOVING. The moving european project - mountain valorisation through interconnectedness and green growth, 2020. Accessed 4 January 2023 <https://www.moving-h2020.eu/>.
- [19] Tadashi Nomoto. Keyword extraction: A modern perspective. *SN Computer Science*, 2023.
- [20] Ollama. Ollama. <https://github.com/ollama/ollama>, 2023.
- [21] Riccardo Orlando, Pere-Lluis Huguet Cabot, Edoardo Barba, and Roberto Navigli. Relik: Retrieve and link, fast and accurate entity linking and relation extraction on an academic budget, 2024.
- [22] Raimundo Real and Juan M Vargas. The probabilistic basis of jaccard’s index of similarity. *Systematic biology*, 45(3):380–385, 1996.
- [23] Ahmad Sakor, Kuldeep Singh, Anery Patel, and Maria-Esther Vidal. Falcon 2.0: An entity and relation linking tool over wikidata. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management, CIKM ’20*, page 3141–3148, New York, NY, USA, 2020. Association for Computing Machinery.
- [24] Brian Schiff. The function of narrative: Toward a narrative psychology of meaning. *Narrative Matters*, 2(1):33–47, 2012.
- [25] Charles Taylor. *Sources of the self: The making of the modern identity*. Harvard University Press, 1992.
- [26] Denny Vrandečić. The rise of wikidata. *IEEE Intelligent Systems*, 28(4):90–95, 2013.
- [27] James V Wertsch and Henry L Roediger. Collective memory: Conceptual foundations and theoretical approaches. *Memory*, 16(3):318–326, 2008.
- [28] Wikidata. SPARQL entity retrieval specifications and examples. https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/queries/examples, 2024.