Evaluation of LLMs on Long-tail Entity Linking in Historical Documents

Marta Boscariol[1], Luana Bulla[2], Lia Draetta[3], Beatrice Fiumanò[4], Emanuele Lenzi[5,7] and Leonardo Piano[6,*]

[1]*Department of Management, University of Turin, Italy*

[2]*Department of Computer Science, University of Catania, Italy*

[3]*Department of Computer Science, University of Turin, Italy*

[4] *Department of Modern Languages, Literatures and Cultures, University of Bologna, Italy*

[5]*Department of Information Engineering (DII), University of Pisa, Italy*

[6]*Department of Mathematics and Computer Science, University of Cagliari, Italy*

[7]*Institute of Information Science and Technologies (ISTI), National Research Council of Italy (CNR), Pisa, Italy*

Abstract

Entity Linking (EL) plays a crucial role in Natural Language Processing (NLP) applications, enabling the disambiguation of entity mentions by linking them to their corresponding entries in a reference knowledge base (KB). Thanks to their deep contextual understanding capabilities, LLMs offer a new perspective to tackle EL, promising better results than traditional methods. Despite the impressive generalization capabilities of LLMs, linking less popular, long-tail entities remains challenging as these entities are often underrepresented in training data and knowledge bases. Furthermore, the long-tail EL task is an understudied problem, and limited studies address it with LLMs. In the present work, we assess the performance of two popular LLMs, GPT and LLama3, in a long-tail entity linking scenario. Using MHERCL v0.1, a manually annotated benchmark of sentences from domain-specific historical texts, we quantitatively compare the performance of LLMs in identifying and linking entities to their corresponding Wikidata entries against that of ReLiK, a state-of-the-art Entity Linking and Relation Extraction framework. Our preliminary experiments reveal that LLMs perform encouragingly well in long-tail EL, indicating that this technology can be a valuable adjunct in filling the gap between head and long-tail EL.

Keywords

Entity linking, Long-tail entities, Large language models, Historical Documents

1. Introduction

Entity Linking (EL) is a fundamental task in Natural Language Processing (NLP) that involves the identification and disambiguation of entity mentions in text, linking them to corresponding entries in a reference Knowledge Base (KB), such as Wikipedia or Wikidata. Accurate EL enhances the understanding of text by connecting unstructured data to structured knowledge, thereby enriching the content with contextual meaning and facilitating more advanced text analytics.

The vast majority of traditional EL approaches typically rely on machine learning [1, 2], some with rule-based approaches [3] and others based on graph optimization [4].

These methods, although effective in many cases, often struggle with ambiguous or obscure mentions, particularly when dealing with long-tail entities [5], i.e. entities that are infrequently mentioned or have limited representation in available KBs. The scarcity of training data and the inherent diversity of long-tail entities make accurate linking a persistent challenge [5].

The advent of Large Language Models (LLMs), such as GPT and Llama, has opened new avenues for EL. Their ability to understand complex language constructs suggests that they could enhance EL

✉ marta.boscariol@unito.it (M. Boscariol); luana.bulla@phd.unict.it (L. Bulla); lia.draetta@unito.it (L. Draetta); beatrice.fiumano@unibo.it (B. Fiumanò); emanuele.lenzi@isti.cnr.it (E. Lenzi); leonardo.piano@unica.it (L. Piano)

ⓘ 0000-0001-7517-6877 (M. Boscariol); 0000-0001-7116-9338 (L. Bulla); 0009-0004-6479-5882 (L. Draetta); 0009-0007-6918-0803 (B. Fiumanò); 0000-0002-9159-8858 (E. Lenzi); 0000-0003-1047-5491 (L. Piano)

performance, especially in contexts where traditional methods falter. LLMs' extensive pre-training on several and diverse corpora allows them to handle a broad range of entities, including those that are less common or not explicitly covered in the training data [6].

To investigate EL performances with long-tail entities and assess the effectiveness of LLMs in this task, the present study addresses two main research questions:

- How does the most reliable state-of-the-art EL tool perform with long-tail entities?
- Are LLMs suitable for long-tail entity linking?

To do so we evaluate the performance of two LLMs (GPT and Llama), in a long-tail entity linking scenario using as benchmark MHERCL v0.1[1] [7], a manually annotated collection of sentences from domain-specific historical texts. By comparing the performance of these LLMs against that of ReLiK [8], a state-of-the-art Entity Linking and Relation Extraction framework, this study aims to shed light on the potential and limitations of LLMs in handling long-tail entity linking in specialized domains.

While long-tail entities are a fairly well-known phenomenon, relatively few researchers have addressed the long-tail EL task. For this reason, our work is part of a research area that is still largely unexplored. Additionally, this study is part of an innovative line of research that leverages LLMs for various knowledge graph-related applications. There is a clear need for further investigation into the potential roles these technologies could play across different contexts.

The present work is organized as follows: in Section 2, we briefly present the related work on entity linking and long-tail entities. Section 3 describes the methodology adopted in the experiment. In Section 4, we introduce the experimental setup, including the dataset we use and the state-of-the-art baseline. Sections 5 and 6 respectively present the results obtained and the final considerations.

2. Related work

Entity linking, the task of associating mentions in text with their corresponding entities in a knowledge base, has been extensively studied. Early approaches relied on heuristic-based methods [9, 10], and among them a prominent system is DbPedia Spotlight [11], which automatically annotates text with DbPedia URIs, combining lexical matching techniques with context-based disambiguation. Significant advancements have been achieved by Neural Entity Linking approaches, that leverage Deep Neural Networks and Languages Models. GENRE [12] employs a sequence-to-sequence approach to autoregressively generate unique entity names. CHOLAN [13] improved EL performance by relying on a modular approach. First, it detects mentions with a BERT transformer, then it retrieves a list of candidate WikiData entities, and finally it employs another BERT model enhanced with local sentence context and Wikipedia entity descriptions to classify and link the mention to the correct entity. The aforementioned approaches involve extracting entity mentions and then linking them to a proper KB, whereas in [14] the author reverses this order by first retrieving candidate entities from the KB and then finding the respective mentions in the text employing a Question Answering strategy. In contrast, *ReLiK* [8] introduces a novel state-of-the-art Entity Linking and Relation Extraction system based on a Retriever-Reader architecture. Its novelty resides in using a single forward pass to link all candidate entities and extract relations, unlike previous methods that require separate passes for each candidate. This strategy permits ReLiK to achieve up to 40x faster inference compared to other methods, while maintaining strong performances. More recently, with the advent of LLMs, several researchers have implemented EL solutions that take advantage of such technology. *ChatEL* [15] is a three-step entity linking framework where, after retrieving a set of candidate entities with *BLINK* [16], an LLM is prompted first to augment the entities mentions with meaningful descriptions to improve disambiguation and then to choose the correct entity. Similarly, the LLMAEL pipeline [17] leverages LLMs as context augmenters for traditional EL models such as GENRE and BLINK, coupling their task specification capabilities with the extensive world knowledge of LLMs. Xin et al. note that this approach also enhances EL performances in long-tail scenarios, as LLMs enrich EL models with additional knowledge on low-frequency entities,

---

facilitating entity identification and linking. Despite these advances, however, the domain of long-tail entity linking remains largely underexplored in current research, as most of the developed systems and datasets are mainly designed to capture head entities [5].

3. LLM-based Entity Linking

Entity linking usually involves two core tasks: (i) Entity Recognition, which detects and extracts the entity mentions from the text, and (ii) Entity Disambiguation, where the entity is correctly linked to its respective Knowledge base entry. As LLMs excel at capturing complex relationships between words, we tackled the EL problem as a sequence-to-sequence translation, jointly performing mention detection and linking with a single model interrogation. Formally, given a sentence S, comprising within it a set of entities E, where each entity is uniquely represented by a unique label (e.g Wikipedia page title), the model needs to identify each entity $e_i \in E$ along with its unique identifier. To accomplish this, we prompted the LLMs to generate a JSON-style output having as a key the textual span of the identified mention and as a value the respective Wikipedia page title. In an autoregressive fashion, the model detects the textual mentions that refer to an entity and consequently translates them into the corresponding unique identifier by probing from its knowledge. To further assist the model, we supplied an example in the prompt, thus following a one-shot approach. The employed prompt is detailed below.

Entity Linking Prompt

```
You are a powerful Entity Linking system.
Given a sentence, identify the key entities and output their exact labels as found on the corresponding Wikipedia pages.
Generate a structured JSON output, formatted as [{"Entities":{"text entity span": "Wikipedia page title"}}].
Here there are some examples:
#
Sentence:"of Rameau was represented in 1735, it was a balletopera Les Indes galantes."
Output:  [{"Entities":{"Rameau":"Jean-Philippe Rameau","Les Indes galantes":"Les Indes galantes"}}]
```

As an alternative identifier, it might be conceivable to exploit the QID, the unique identifier of Wikidata entities. However, through a preliminary experiment, we noticed that LLMs tended to fictionalize QIDs. In that experiment, GPT 3.5 achieved a precision of less than 1%. We hypothesize that this behavior is caused by the fact that QIDs mainly consist of numbers and since they don't follow linguistic patterns, LLMs, which are trained primarily on text don't intuitively know how to generate them accurately. As a result, LLMs end up generating plausible-sounding yet fictional QIDs based on learned patterns. For the aforementioned reasons, we decided to exploit the Wikipedia page title as the unique identifier. For clarity, we also specify that we employ the same strategy and prompt for all the compared LLMs, which are detailed in the following section.

4. Experimental setup

Models. For our study, we harness two advanced LLMs, namely GPT 3.5 and LLama 3 [18], both in their instruct versions. Specifically, our experiments are conducted using OpenAI's `GPT-3.5-turbo-instruct`[2], Meta `LLama-3-8B-instruct`[3] and `Llama-3-70B`[4]. GPT 3.5 Turbo is a cost-effective, cutting-edge tool ensuring deep contextual understanding, heightened accuracy and faster processing speed compared to other GPT models. LLama 3, available in configurations with 8 billion and 70 billion parameters, is an open-source, highly versatile model that offers state-of-the-art performances in a wide variety of NLP tasks, outperforming GPT models in different benchmarks and having longer context window compared to GPT-3.5 Turbo.

Dataset. The performance of LLMs is evaluated on the Musical Heritage Historical named Entities Recognition, Classification and Linking (MHERCL) benchmark. MHERCL v0.1 consists of English-

---

language sentences extracted from the *Periodicals* module of the Polifonia Textual Corpus[5] (PTC), a diachronic corpus covering the domain of Musical Heritage. As part of the PTC creation, Optical Character Recognition (OCR) was leveraged to extract text from scans of historical documents. Since OCR on historical documents can be particularly challenging and prone to errors due to factors like degraded image quality or archaic fonts, this process inevitably introduced some noise into the dataset. Around 930 sentences were extracted from the PTC to create the MHERCL benchmark. Each sentence was manually annotated with EL information, including entity type and Wikidata QID. In this work, we use MHERCL v0.1.2, an expanded version of the benchmark that includes 928 sentences, 969 unique named entity mentions (NE) identified by a QID, and 59 different NE types. Table 1 provides a synthetic overview of the dataset statistics. Entities that were not assigned a QID, mainly due to the lack of a corresponding Wikidata entry, were assigned a NIL label and are excluded from the dataset statistics. Since the *Periodicals* module of the PTC consists of music-specialised documents published between 1823 and 1900, the MHERCL dataset features a high concentration of niche, domain-specific and historical knowledge, serving as a robust benchmark for assessing EL performances in long-tail scenarios.

Dataset	Lang.	Sent.	Tokens	Unique NE	Types
MHERCL v0.1.2	EN	928	28.478	966	59

Table 1
MHERCL v0.1.2 Benchmark statistics

Baseline. As our baseline, we leverage ReLiK, a state-of-the-art framework for Entity Linking and Relation Extraction, on the MHERCL dataset. Based on a retriever-reader architecture, ReLiK outperforms its competitors in both in-domain and out-of-domain settings, achieving better results in terms of performance, inference speed, and flexibility. ReLiK is available in three versions: small, base, and large. In our study, we leverage ReLiK-base to identify and link entity mentions within MHERCL sentences to their corresponding Wikidata entries. ReLiK links entities to a knowledge base other than WikiData, providing Wikipedia page IDs instead of WikiData QIDs. To map the extracted entities to their corresponding WikiData entries, we queried ReLiK's reference knowledge base, KILT [19]. KILT, derived from a Wikipedia dump from August 1, 2019, allowed us to retrieve WikiData IDs using either Wikipedia page IDs or entity titles.

Evaluation. To evaluate and compare the performance of the selected models, we employ confusion matrix metrics such as precision, recall, and F1-score, which are formally defined as:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

In computing these scores, we count a True Positive (TP) when the model's prediction matches the ground truth, a False Positive (FP) when the model's prediction does not match annotations in the ground truth, and a False Negative (FN) when the model is not able to identify and link the entity. Entities labeled as NIL in the MHERCL benchmark are excluded from our experiments because they generally lack corresponding Wikidata or Wikipedia entries. We further specify that, in the case of ReLiK's results, we evaluated the match between the QID identified by ReLiK and the corresponding QID in the ground truth. In contrast, for LLMs we assessed the correct match between the predicted Wikipedia page title and the Wikipedia title retrieved using the baseline's QID.

---

5. Results

This section outlines the quantitative results of our preliminary study in historical long-tail Entity Linking with Large Language Models. Table 2 highlights the comparison between LLMs against ReLiK, over the entire dataset, with no differentiation in the distribution of entities. This comparison shows that ReLiK is highly accurate, generating a low number of false positives, reaching a precision of 72.8%. Still, it struggles to find an adequate amount of entities in such a niche domain, retrieving the 45% of annotated entities as evidenced by the recall. Indeed, exception made for LLama3-8b, the LLMs recovered a higher number of entities, where LLama3 in the 70b configuration reached a recall score of 60.3% exceeding the state-of-the-art ReLiK by about 15%. Given the high recall of entities made by LLMs, we hypothesize that they could serve as entity retrievers or augment the retrieval of existing EL retrievers and we believe that this aspect may provide insights for future studies and investigations from the scientific community. Regardless of satisfactory recall results, the numbers dropped in the precision instance, this is because the LLMs always tended to generate some text and find fictional entities that were not annotated in the dataset, raising the number of false positives. Also, our evaluation metrics are based on exact matching between the predicted Wikipedia label and the real one, so even a single incorrectly generated character causes the prediction to be considered incorrect. Overall, It is worth noting how the obtained results highlight the potential of LLMs in EL, as even though they are not specifically trained on the EL task, they achieved competitive results with respect to the state of the art, in an out-of-domain comparison. They even exceeded it when it comes to long-tail entity retrieval.

Table 2
Comparison between LLMs and Relik

Model	Precision(%)	Recall(%)	F1(%)
ReLiK	**72.8**	45.7	**56.1**
GPT 3.5	48.6	58.8	53.2
LLama3-70b	47.3	**60.3**	53.0
LLama3-8b	34.9	40.1	37.3

Although the MHERCL benchmark is a historical, domain-specific dataset with a high number of niche and less popular entities, we conducted an additional analysis to highlight better the models' performance when varying the entities' popularity. As a measure of popularity, we leveraged the number of Wikidata triples associated with each entity, as also done in [20].
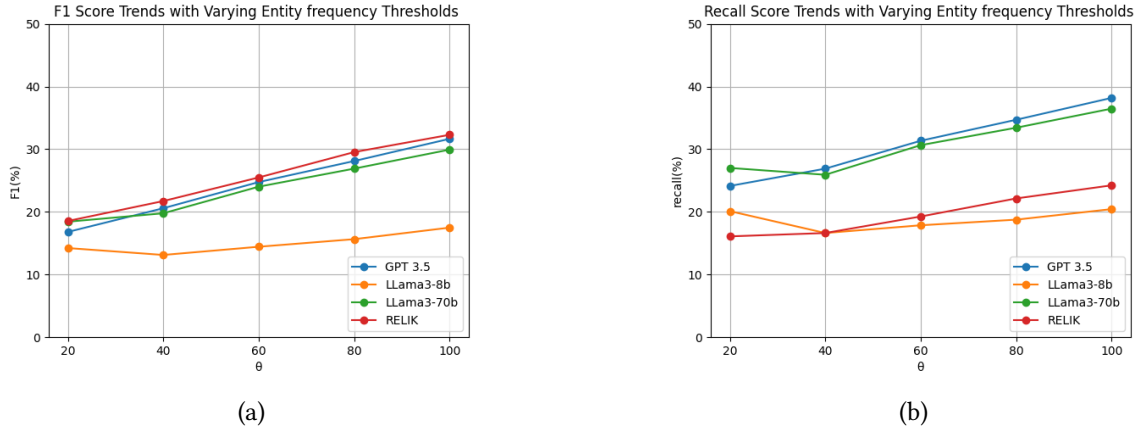

(a)   (b)

Figure 1: Measurement of the Entity Linking F1 and Recall score Across Different Entity Occurrence Thresholds for all the employed models

The plot depicted in Figure 1a, reports the variation of the F1 score in EL at the variation of a threshold θ, which has the role of differentiating real and predicted entities based on their notoriety defined by

the number of Wikidata triples being associated to them. For example, a threshold θ of 20, takes into account all entities that have at most 20 triples associated with them. The plot demonstrates that higher entity frequency thresholds θ generally lead to better EL performance for all models, likely because higher thresholds focus on more frequent, well-represented entities that are easier to disambiguate and link correctly. LLama3-70 achieves the highest f1 on par with ReLiK in linking very rare entities, having a threshold $\theta = 20$. The plot in Figure 1b instead, highlights the recall fluctuation. GPT-3.5 and Llama 3-70b models perform better overall, with increasing recall scores as the threshold increases. ReLiK, despite being a specialized tool stays below the two largest LLMs and performs poorly with infrequent entities, obtaining the lowest recall score in the case of $\theta = 20$. Llama 3-8b instead, demonstrates consistently lower recall and does not show significant gains with increasing entity frequency. The conducted analyses clearly show that the entity linking of long-tail entities is still an open challenge, as one of the most performant state-of-the-art tools was only able to retrieve $\approx 15\%$ of the annotated entities when they were less known and possessed a low frequency index. On the other hand, LLMs, at least in the larger configurations retrieved a higher number of entities, but the numbers remained unsatisfactory with a recall below 30% and an F1 of $\approx 19\%$ referring to the less popular entities.

Qualitative evaluation. For the sake of comprehensiveness and further interpretation of the results, we conducted a brief qualitative analysis. Upon closer examination, we observed that a small number of entities were not correctly disambiguated by the models, due to spelling errors introduced by the OCR on the original documents. Noisy text elements are common when working with digitalized texts, especially digitalized historical documents. While human annotators were able to easily detect these mistakes and accurately identify the correct entities, many models struggled to move beyond the surface-level text. Specifically, when limited semantic context was available around the entity, both the baseline models and the LLMs struggled to accurately perform EL. For instance, given the sentence

```
'Mr. Mocre is the adaptor of words to this composition, which is a tirana,
arranged by Mr. Bishop.'
```

none of the models were able to associate the form *Mocre* with Thomas Moore (Q315346). On the other hand, when provided with sufficient contextual information about the entities, LLMs were more likely to identify the correct entity even in the presence of lexical errors. For example, in the sentence

```
'One man may lived, who ean read the heart, and whose power was not: based upon,
his own experience but if so, we may well call William Shakspeare superhuman,
THenee it was that whiffe i m Rossint's 'Barber of Seville,' ar Cimarosa's
'Seeret Marriage'
```

despite the inherent difficulty due to the OCR mistakes, both GPT and LLama80b correctly associated 'Rossint' with the composer Gioachino Rossini and 'Seeret Marriage' with Domenico Cimarosa's work *The Secret Marriage.*

Nevertheless, when it comes to sentences that include less popular entities, even with appropriate context, LLMs may struggle to properly disambiguate the involved entities. For example, when encountering the phrase 'Teatro Santo Augustino in Genoa', which should be linked to Teatro Sant'Agostino (Q19060499), both GPT 3.5 Turbo and Llama-70B incorrectly linked the entity to the more renowned Teatro Carlo Felice in Genoa.

6. Conclusion and future works

In conclusion, this study highlights the potential of large language models (LLMs), such as GPT and Llama, for improving entity linking, particularly in challenging long-tail scenarios. While state-of-the-art systems like ReLiK perform well on frequent entities, LLMs show a significant advantage in identifying and linking less common, domain-specific entities, as evidenced by their higher recall scores. Despite lower precision due to occasional over-generation of entities, LLMs demonstrate the potential

to recover more long-tail entities compared to ReLiK. This suggests that LLMs can serve as valuable tools in bridging the gap between frequent and infrequent entities in historical and domain-specific contexts. Furthermore, this study represents an early, exploratory effort to understand the efficacy of LLMs in the long-tail entity linking (EL) scenario, wherein we employed and tested relatively simple, vanilla prompt-based approaches. Thus suggesting that LLMs, even with their base unmodified form, possess inherent advantages over traditional entity-linking systems. However, there remains significant potential for further refinement. More emphasis should be placed on optimizing the balance between recall and precision. While recall is an important metric, especially for long-tail entities, precision must not be overlooked. Thus, future work should focus on developing more sophisticated prompting strategies or hybrid systems. Possible investigations include In-Context Learning (ICL) techniques, to better tailor LLMs to the task of entity linking or Knowledge Injection, to augment the LLMs' knowledge and their contextual understanding. Such methods could potentially mitigate the over-generation issue, while enhancing their accuracy in identifying and linking entities in more narrow contexts.

References

[1] D. Banerjee, D. Chaudhuri, M. Dubey, J. Lehmann, Pnel: Pointer network based end-to-end entity linking over knowledge graphs, in: The Semantic Web–ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part I 19, Springer, 2020, pp. 21–38.

[2] E. Boros, E. L. Pontes, L. A. Cabrera-Diego, A. Hamdi, J. G. Moreno, N. Sidère, A. Doucet, Robust named entity recognition and linking on historical multilingual documents, in: Conference and Labs of the Evaluation Forum (CLEF 2020), volume 2696, CEUR-WS Working Notes, 2020, pp. 1–17.

[3] A. Sakor, K. Singh, A. Patel, M.-E. Vidal, Falcon 2.0: An entity and relation linking tool over wikidata, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 3141–3148.

[4] M. Klang, P. Nugues, Hedwig: A named entity linker, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, 2020, pp. 4501–4508.

[5] F. Ilievski, P. Vossen, S. Schlobach, Systematic study of long tail phenomena in entity linking, in: E. M. Bender, L. Derczynski, P. Isabelle (Eds.), Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 664–674. URL: https://aclanthology.org/C18-1056.

[6] B. Li, G. Fang, Y. Yang, Q. Wang, W. Ye, W. Zhao, S. Zhang, Evaluating chatgpt's information extraction capabilities: An assessment of performance, explainability, calibration, and faithfulness [arxiv: 2304.11633 [cs]], arXiv preprint arXiv:2304.11633 (2023).

[7] A. Graciotti, Knowledge extraction from multilingual and historical texts for advanced question answering, in: C. d'Amato, J. Z. Pan (Eds.), Proceedings of the Doctoral Consortium at ISWC 2023 co-located with 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 7, 2023, volume 3678 of *CEUR Workshop Proceedings*, 2023.

[8] R. Orlando, P.-L. Huguet-Cabot, E. Barba, R. Navigli, Relik: Retrieve and link, fast and accurate entity linking and relation extraction on an academic budget, arXiv preprint arXiv:2408.00103 (2024).

[9] K. Nguyen, R. Ichise, Heuristic-based configuration learning for linked data instance matching, in: Semantic Technology: 5th Joint International Conference, JIST 2015, Yichang, China, November 11-13, 2015, Revised Selected Papers 5, Springer, 2016, pp. 56–72.

[10] G. Zheng, M. Liu, B. Liu, Collective entity linking based on dbpedia, in: Knowledge Graph and Semantic Computing. Language, Knowledge, and Intelligence: Second China Conference, CCKS 2017, Chengdu, China, August 26–29, 2017, Revised Selected Papers 2, Springer, 2017, pp. 66–79.

[11] P. N. Mendes, M. Jakob, A. García-Silva, C. Bizer, Dbpedia spotlight: shedding light on the web of documents, in: Proceedings of the 7th international conference on semantic systems, 2011, pp. 1–8.

[12] N. De Cao, G. Izacard, S. Riedel, F. Petroni, Autoregressive entity retrieval, arXiv preprint arXiv:2010.00904 (2020).

[13] M. P. K. Ravi, K. Singh, I. O. Mulang, S. Shekarpour, J. Hoffart, J. Lehmann, Cholan: A modular approach for neural entity linking on wikipedia and wikidata, arXiv preprint arXiv:2101.09969 (2021).

[14] W. Zhang, W. Hua, K. Stratos, Entqa: Entity linking as question answering, arXiv preprint arXiv:2110.02369 (2021).

[15] Y. Ding, Q. Zeng, T. Weninger, Chatel: Entity linking with chatbots, arXiv preprint arXiv:2402.14858 (2024).

[16] L. Wu, F. Petroni, M. Josifoski, S. Riedel, L. Zettlemoyer, Scalable zero-shot entity linking with dense entity retrieval, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6397–6407. URL: https://aclanthology.org/2020.emnlp-main.519. doi:10.18653/v1/2020.emnlp-main.519.

[17] A. Xin, Y. Qi, Z. Yao, F. Zhu, K. Zeng, X. Bin, L. Hou, J. Li, Llmael: Large language models are good context augmenters for entity linking, 2024. URL: https://arxiv.org/abs/2407.04020. arXiv:2407.04020.

[18] Llama Team, AI @ Meta, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[19] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, et al., Kilt: a benchmark for knowledge intensive language tasks, arXiv preprint arXiv:2009.02252 (2020).

[20] L. Chen, S. Razniewski, G. Weikum, Knowledge base completion for long-tail entities, arXiv preprint arXiv:2306.17472 (2023).