Enriching Metadata with LLMs and Knowledge Graphs: Case Finnish Named Entity Linking

Rafael Leal[1,*], Annastiina Ahola[1] and Eero Hyvönen[1,2]

[1]*Aalto University, Department of Computer Science, https://seco.cs.aalto.fi/*
[2]*University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG)*

Abstract

This paper presents work on using Large Language Models (LLM) to disambiguate Named Entity Linking candidates, which is meant for enriching the metadata of textual documents by linking them to Knowledge Graphs. We propose a zero-shot classification method that has similarities with Retrieval-Augmented Generation (RAG), and discuss 1) a prototype web service and 2) a user interface on top of it that allows for human intervention when making final disambiguation decisions, especially when this cannot be reliably carried out in automatic fashion due to errors and hallucinations of LLM-based tools. The focus of this work is on Finnish texts, so our methods take into account the particularities of this highly inflectional language and the resources available for processing it. The paper presents promising preliminary evaluation results of the system, suggesting feasibility of the methods and tools presented: our named entity lemmatizer achieved an accuracy of 96.5% on our test dataset, and a local LLM of the Llama family was able to find the correct linking candidate in 16 out of 17 examples. GPT-4 achieved 100% accuracy in linking using both standard text and YAML format.

Keywords

Large language models, Named Entity Recognition, Named Entity Linking, Linked data, Knowledge organization system, Knowledge graph,

1. Introduction and Motivation

Much of the data that could be used in Digital Humanities (DH) research is available only in unstructured textual form. Information extraction is then needed for creating metadata based on Knowledge Organization Systems (KOS) and Knowledge Graphs (KG) (Martinez-Rodriguez, Hogan, and Lopez-Arevalo 2020), publishing Linked Data Services, and building applications on top of them, such as the Sampo systems (Hyvönen 2022). For example, in our work on publishing the plenary session speeches of the Parliament of Finland as Linked Open Data (LOD), the speeches had to be linked to various domain-specific ontologies based on named entities (people, places, organizations, etc.), keyword resources, and a library classification system (Tamper et al. 2022). A fundamental task here is Named Entity Recognition (NER) and Linking (NEL). This paper addresses the question of how Large Language Models (LLM) can be

Digital Humanities in the Nordic and Baltic Countries 2025

*Corresponding author.
rafael.leal@aalto.fi (R. Leal); annastiina.ahola@aalto.fi (A. Ahola); eero.hyvonen@aalto.fi (E. Hyvönen)
https://seco.cs.aalto.fi/u/eahyvone (E. Hyvönen)
0000-0001-7266-2036 (R. Leal); 0009-0008-6369-4712 (A. Ahola); 0000-0003-1695-5840 (E. Hyvönen)

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

exploited for the task where semantic disambiguation is a key challenge. This work is focused on Finnish texts, and we discuss some of the pitfalls that incur when carrying out natural language processing in this language.

It is also important to notice that one of our aims is to be aware of the computational power needed to run the system. In many cases, we chose much smaller, specialized local models when a call to an external LLM could be carried out. Moreover, we also strive to use open tools wherever possible.

2. Related works

Traditionally, Named Entity Recognition and Named Entity Liking have been separated into different tasks, although newer works, capitalizing on the breakthroughs of deep neural networks, focus on end-to-end linking. For example, Ayoola et al. (2022) performs entity detection and disambiguation in a single forward pass, while Logeswaran et al. (2019) created a zero-shot linking system focused on addressing domain-shifting, without the need for gazeteers.

With the advent of LLMs, the field of Retrieval-Augmented Generation (Lewis et al. 2020) expanded dramatically, as summarized in Fan et al. (2024). The interface between LLMs an knowledge graphs can be considered a subdomain of RAG, and there are many recent works in this field: Baek, Aji, and Saffari (2023) created a LLM-based framework to answer questions by verbalizing triples related to entities in knowledge graphs, while Edge et al. (2024) used abstractive summarization over an entire corpus to find answers to global questions

Regarding works about the Finnish language, Mäkelä (2014) described a web-based tool for annotating texts based on linked data, which included named entities. Luoma et al. (2020) released a corpus and tool for Finnish Named Entity Recognition, expanding on Ruokolainen et al. (2020). This work has been used for example for pseudonymization of court documents (Oksanen et al. 2019) as well as studying Parliamentary data (Tamper et al. 2022). However, as far as we are aware, there are no previous works using neural network techniques for entity linking in Finnish.

3. Disambiguation and Linking

NER in Finnish is a practical task that has been well-served since the publication of the command-line tool by the TurkuNLP group in 2020 (Luoma et al. 2020). However, the additional, more difficult task of disambiguating and linking entities (NEL) to external KOS and knowledge graphs (KG) has not been as prominent, despite the advent of Large Language Models (LLMs). In this paper, we utilize a classification method that bears similarities with RAG in order to disambiguate and link entities in Finnish texts to linked data resources.

Our approach follows the traditional entity linking procedure, divided into three steps: 1) entity recognition; 2) candidate generation; 3) disambiguation and linking. Our system recognizes entities via third-party tools, such as the aforementioned NER tool. Although LLMs could be used for this purpose, their ratio of computing power demands to accuracy can be significantly worse, and their idiosyncrasies can make it more difficult to extract the entities themselves. The entities are afterwards lemmatized (changed to their basic forms) and lexically matched to find suitable candidates. There are around 15 nominal cases in Finnish, which makes it a significantly harder language to lemmatize than more analytic ones, for example English, which tend to use separate words to indicate case. Since word-by-word lemmatization does not produce satisfactory results for Finnish named entities, we fine-tuned `Finnish-NLP/t5-small-nl24-finnish`[1], a T5 generative model containing around 260 million parameters, with a dataset of Wikipedia internal links of our creation. The training set contains around 1 million pairs of named entity surface forms (alongside their context) and their basic form (based on their page names). This dataset will be released as open data.

The candidates are generated from a local database, containing Wikipedia articles and related Wikidata, which have long been used as primary targets for named entity linking (Mihalcea and Csomai 2007). The generation is done via lexical matching, which means that word forms, rather than their meaning, are used to find similarities. This technique might be more fragile than alternatives based on vectorization, since it uses word forms themselves for the generation, instead of context. It also requires a wealth of alternative labels to be included in the databases, so that entities can also be found via surrogate names. However, it allows for easy integration of any number of databases and knowledge graphs, since it does not require pre-processing or fine-tuning, and dispenses with tracking changes to them.

For many named entities, the number of plausible candidates can be heuristically shrunk to one, which bypasses the need for further processing. Otherwise, the candidates are presented to the LLM, which is tasked with deciding which one is the most suitable, based on the context in which the entity appears in the text. This step has some characteristics in common with RAG, since both use external retrieval to enhance prompting and capitalize on the emergent capabilities of LLMs to learn in-context (Chan et al. 2022). However, here the LLM works as a zero-shot classifier rather than a generative model: the information presented to the LLM represent a strict narrowing of generative output rather than contextual information to draw upon. This restriction lessens the tendency of LLMs to hallucinate. Furthermore, asking the model to spell out the reasoning behind its choices is case known to improve generation (Kojima et al. 2022).

Our aim is to link to candidates in any number of databases. In order to achieve this KG-agnostic status, the information related to the candidates extracted from the knowledge graphs should be presented to the LLM in an appropriate format. Standard text descriptions are better suited than knowledge graphs for LLMs to reason upon, which is expected due to the nature of LLM pre-training. However, extended textual descriptions are not common in knowledge bases, so as an alternative setup we retrieve Wikidata properties for the candidates and format them as YAML. The property labels are in Finnish when possible, or in English as a backup. This format was chosen for the ease of conversion from RDF graphs and its minimal markup, which translates into fewer LLM tokens.

4. A Tool for Automatic Annotation

The purpose of our research is to create not only a strong baseline for automatic annotation of Finnish documents but also a user interface that supports its application. As such, the entity

[1] https://huggingface.co/Finnish-NLP/t5-small-nl24-finnish

linking tool is being integrated into a front-end interface that could enable users to seamlessly revise and correct the named entity links found in a document and then save the metadata for further use in DH research and applications. Such a tool can be used in cases where the annotations are critically important, such as in dealing with legal documents or parliamentary data[2] (Hyvönen et al. 2024). The user would then be able to upload the text to the tool, review and edit entity links proposed by the tool, and the save the corrected metadata in a fashion similar to the pseudonymization tool ANOPPI (Oksanen et al. 2019), previously created by our research group.

5. Results

5.1. Named entity lemmatization

Our named entity lemmatizer obtained an accuracy of 96.5% on 10K test examples from the Wikipedia internal links dataset. Since many characters have to be added to the model's tokenizer in order to correctly process foreign-language entities, we also fine-tuned a multilingual version of T5[3], which include such characters during training, but this model only achieved an accuracy of around 83%. It seems that the multilingual version does not incorporate Finnish grammatical rules well enough for it to succeed in this task.

5.2. Entity disambiguation

Two LLM models were tested for entity disambiguation: Llama-3-8B-Instruct(Touvron et al. 2023) and GPT-4(OpenAI et al. 2024). A total of 17 disambiguation examples were used for this task. These examples were automatically extracted from a Finnish Wikinews dataset that we have labelled, and are all related to Wikipedia disambiguation pages. This test sample includes, among others, locations (Gothenburg, Odessa (Texas), Kaduna), organizations (Esso, Citroën), and persons (John Roberts). There is an average of 5.7 candidates for each entity in this test dataset.

Two different kinds of candidate descriptions were tested: 1) Standard text, using the introductory part of the respective Wikipedia articles for each candidate, and 2) Wikidata contents formatted as YAML, as mentioned above. Only 5 examples were used in this task.

Llama-3-8B-Instruct correctly identified 16 out of 17 cases using standard text, but it failed in all 5 cases using YAML. The model understands the task and returns one of the candidates, but it cannot identify the correct one. GPT-4, on the other hand, obtained an accuracy of 100% in both tasks.

6. Discussion

Our main objective in working on this project is to increase the digital presence of Finnish, making the existing metadata-processing tools for this language more robust and full-featured.

[2] Our group has released, among others, the LawSampo (https://lakisampo.fi/)and ParliamentSampo (https://parlamenttisampo.fi/) data services and semantic portals which use linked data.

[3] google/mt5-small

However, working with a language which is so peripheral even in an European context is challenging, which can be seen in practice in the absence of tools and resources such as open datasets.

This paper describes an entity linking system in Finnish, which capitalizes on existing tools and models in order to extract named entities, generate candidates for them and choose the best candidate for linking. It showed that existing LLMs are capable of dealing with Finnish text satisfactorily, and the most powerful ones are even able to parse predicates in YAML with full accuracy. These results suggest the feasibility of the methods and tools presented here.

In the future, we will test larger local LLMs, such as Llama-3.1-70B-Instruct, to probe their capabilities of disambiguating candidates in YAML. This would allow us to utilize knowledge graphs that do not store standard text descriptions while avoid committing to a single LLM. Furthermore, our front-end interface for editing links will be finalized and integrated with our system. It will be open-sourced in its entirety, alongside our dataset of fully-linked Wikinews texts.

References

Ayoola, Tom, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. "ReFinED: An Efficient Zero-shot-capable Approach to End-to-End Entity Linking." In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track,* 209–220. NAACL-HLT 2022. Hybrid: Seattle, Washington + Online: Association for Computational Linguistics, July. https://doi.org/10.18653/v1/2022.naacl-industry.24.

Baek, Jinheon, Alham Aji, and Amir Saffari. 2023. "Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering." In *Proceedings of the First Workshop on Matching From Unstructured and Structured Data (MATCHING 2023),* edited by Estevam Hruschka, Tom Mitchell, Sajjadur Rahman, Dunja Mladenić, and Marko Grobelnik, 70–98. MATCHING 2023. Toronto, ON, Canada: Association for Computational Linguistics, July. https://doi.org/10.18653/v1/2023.matching-1.7.

Chan, Stephanie, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. "Data Distributional Properties Drive Emergent In-Context Learning in Transformers." In *Advances in Neural Information Processing Systems,* edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, 35:18878–18891. Curran Associates, Inc.

Edge, Darren, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. "From Local to Global: A Graph RAG Approach to Query-Focused Summarization." Pre-published, April 24, 2024. https://doi.org/10.48550/arXiv.2404.16130. arXiv: 2404.16130 [cs].

Fan, Wenqi, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. "A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models." In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining,* 6491–6501. KDD '24. New York, NY, USA: Association for Computing Machinery, August 24, 2024. ISBN: 9798400704901. https://doi.org/10.1145/3637528.3671470.

Hyvönen, Eero. 2022. "Digital Humanities on the Semantic Web: Sampo Model and Portal Series." *Semantic Web – Interoperability, Usability, Applicability* 14 (4): 729–744. https://doi.org/10.3233/SW-190386.

Hyvönen, Eero, Laura Sinikallio, Petri Leskinen, Senka Drobac, Rafael Leal, Matti La Mela, Jouni Tuominen, Henna Poikkimäki, and Heikki Rantala. 2024. "Publishing and Using Parliamentary Linked Data on the Semantic Web: ParliamentSampo System for Parliament of Finland." Accepted. https://www.semantic-web-journal.net/system/files/swj3605.pdf.

Kojima, Takeshi, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. "Large Language Models Are Zero-Shot Reasoners." In *Advances in Neural Information Processing Systems,* edited by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, 35:22199–22213. Curran Associates, Inc.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. 2020. "Retrieval-augmented generation for knowledge-intensive NLP tasks." In *Proceedings of the 34th International Conference on Neural Information Processing Systems.* NIPS '20. , Vancouver, BC, Canada, Curran Associates Inc. ISBN: 9781713829546.

Logeswaran, Lajanugen, Ming-Wei Chang, Kenton Lee, Kristina Toutanova, Jacob Devlin, and Honglak Lee. 2019. "Zero-Shot Entity Linking by Reading Entity Descriptions." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* edited by Anna Korhonen, David Traum, and Lluís Màrquez, 3449–3460. ACL 2019. Florence, Italy: Association for Computational Linguistics, July. https://doi.org/10.18653/v1/P19-1335.

Luoma, Jouni, Miika Oinonen, Maria Pyykönen, Veronika Laippala, and Sampo Pyysalo. 2020. "A Broad-coverage Corpus for Finnish Named Entity Recognition." In *Proceedings of the 12th Language Resources and Evaluation Conference,* 4615–4624. LREC 2020. Marseille, France: European Language Resources Association, May. ISBN: 979-10-95546-34-4, accessed November 30, 2021. https://aclanthology.org/2020.lrec-1.567.

Mäkelä, Eetu. 2014. "Combining a REST Lexical Analysis Web Service with SPARQL for Mashup Semantic Annotation from Text." In *The Semantic Web: ESWC 2014 Satellite Events,* edited by Valentina Presutti, Eva Blomqvist, Raphael Troncy, Harald Sack, Ioannis Papadakis, and Anna Tordai, 424–428. Cham: Springer International Publishing. ISBN: 978-3-319-11955-7. https://doi.org/10.1007/978-3-319-11955-7_60.

Martinez-Rodriguez, Jose L., Aidan Hogan, and Ivan Lopez-Arevalo. 2020. "Information Extraction Meets the Semantic Web: A Survey." *Semantic Web – Interoperability, Usability, Applicability* 11 (2): 255–335.

Mihalcea, Rada, and Andras Csomai. 2007. “Wikify!: Linking Documents to Encyclopedic Knowledge.” In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management,* 233–242. CIKM07: Conference on Information and Knowledge Management. Lisbon Portugal: ACM, November 6, 2007. ISBN: 978-1-59593-803-9. https://doi.org/10.1145/1321440.1321475.

Oksanen, Arttu, Minna Tamper, Jouni Tuominen, Aki Hietanen, and Eero Hyvönen. 2019. “Anoppi: A Pseudonymization Service for Finnish Court Documents.” In *Legal Knowledge and Information Systems. JURIX 2019: The Thirty-Second Annual Conference,* edited by M. Araszkiewicz and V. Rodríguez-Doncel, 251–254. IOS Press, December. ISBN: 978-1-64368-048-4. https://doi.org/10.3233/FAIA190335.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. 2024. *GPT-4 Technical Report.* arXiv: 2303.08774 [cs.CL].

Ruokolainen, Teemu, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2020. “A Finnish News Corpus for Named Entity Recognition.” *Language Resources and Evaluation* 54, no. 1 (March 1, 2020): 247–272. ISSN: 1574-0218. https://doi.org/10.1007/s10579-019-09471-7.

Tamper, Minna, Rafael Leal, Laura Sinikallio, Petri Leskinen, Jouni Tuominen, and Eero Hyvönen. 2022. “Extracting Knowledge from Parliamentary Debates for Studying Political Culture and Language.” In *Proceedings of the 1st International Workshop on Knowledge Graph Generation From Text and the 1st International Workshop on Modular Knowledge co-located with 19th Extended Semantic Conference (ESWC 2022),* edited by Sanju Tiwari, Nandana Mihindukulasooriya, Francesco Osborne, Dimitris Kontokostas, Jennifer D’Souza, and Mayank Kejriwal, 3184:70–79. International Workshop on Knowledge Graph Generation from Text (TEXT2KG 2022). CEUR WS, May. http://ceur-ws.org/Vol-3184/TEXT2KG_Paper_5.pdf.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. 2023. *LLaMA: Open and Efficient Foundation Language Models.* arXiv: 2302.13971 [cs.CL].