”

Using LLMs for Enriching Metadata with Links to KOS and Knowledge Graphs: Case Finnish Named Entity Linking

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

[1]*Semantic Computing Research Group (SeCo), Department of Computer Science, Aalto University, Finland*
[2]*University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG), Finland*

Abstract

This paper presents work on using Large Language Models (LLM) for disambiguating Named Entity Linking candidates, which is meant for enriching the metadata of textual documents by linking them to Knowledge Organization Systems, a.k.a domain ontologies, and Knowledge Graphs. We propose a zero-shot classification method that has similarities with Retrieval-Augmented Generation (RAG), and discuss an under-development prototype tool that allows for human intervention when making final disambiguation decisions, especially when this cannot be reliably carried out in automatic fashion. The focus of this work is on Finnish texts, so our methods must take into account the particularities of this language and the resources available for processing it.

1. Enriching Metadata of Texts by Data Linking

Much of the data that could be used in Digital Humanities 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) [1], publishing Linked Data Services, and building applications on top of them, such as the Sampo systems [2]. 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 [3]. 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 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 Finnish.

CFP, NKOS Workshop 2024
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)

2. Disambiguation and Linking

NER in Finnish is a task that has been well-served since the publication of a broad-coverage corpus by the TurkuNLP group in 2020 alongside a command line-based tool [4]. However, the additional, more difficult task of disambiguating and linking entities (NEL) to external KOS and knowledge graphs (KG) has not advanced much in recent years, despite the advent of Large Language Models (LLMs). In this paper, we propose a classification method that bears similarities with Retrieval-Augmented Generation (RAG) [5] and its corresponding tool 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: entity recognition; candidate generation; disambiguation and linking. Our system recognizes entities via external, BERT-based [6] classification tools. Although LLMs could be used for this purpose, their ratio of computing power to accuracy can be significantly worse, and they tend to not report starting and ending indices accurately. In the current system, the entities are afterwards lemmatized (changed to their basic forms) and lexically matched to find suitable candidates. Lexical matching means that word forms, rather than their meaning, are used to find similarities. This is, however, a fragile method, since the around 15 nominal cases in Finnish make it a significantly harder language to lemmatize than more analytic ones, for example English, which tend to use words such as prepositions to indicate case. Moreover, this technique relies on alternative names and spellings being listed in the databases, alongside canonical ones, as labels.

Consequently, we are investigating vector-based alternatives for candidate generation, although vector search may be incompatible with the aim of using any number of different databases. Efficient vectorization implies pre-processing of the targets: in order to find matching candidates in a vector space, using for example using cosine similarity, it is necessary to process and compare all possible targets. Moreover, performant models should be fine-tuned, but there do not exist suitable open datasets for NEL fine-tuning in Finnish. Lexical matching is thus our current method for candidate generation.

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 [7]. 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, although do not eliminate it, since the reasoning behind the conclusion is not controlled by the prompt. Furthermore, although more advanced LLMs tend to respond consistently, there is no guarantee that the candidate that emerge as product of the reasoning will be the same one highlighted by the model: it might for example conclude that candidate "B" is the best yet answer with a contender, such as "C". Asking the model to spell out the reasoning behind its choices is in any case known to improve generation [8].

As stated previously, 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 has to be presented to the LLM in an appropriate format. Wikipedia-like descriptions are better suited than for example Wikidata content for LLMs to reason upon, which is expected due to the nature of LLM pre-training. However, extended descriptions are not common in knowledge bases, so alternative formats must be found. We are researching low-markup format such as YAML – in order to avoid computational overhead – and prompting techniques that intend to capitalize on the characteristics of knowledge graphs. This is one of the main challenges that must be overcome for this project.

Also, the accuracy of the candidate classification task is highly dependent on the LLM used, since the capacity of reasoning and learning in-context is model-dependent. According to our initial tests, OpenAI's GPT-4 seems to be the most accurate in pointing the correct answer using both Wikipedia descriptions and Wikidata in YAML format, although the lack of pre-built datasets has made evaluation harder. However, we strongly favour the usage of open-source tools in order to restrengthen the practice of open science, enable repeatability and avoid lock-ins. This rules out many among the most advanced LLMs available nowadays, such as the aforementioned GPT family or Gemini by Google.

3. A Tool for Automatic Annotation

The purpose of our research is not only to create a strong baseline for automatic annotation of Finnish documents but also to create an open-source tool that supports its application. This entity linking prototype, still under development, is command line-based, but can be integrated into a front-end interface that could enable users to seamlessly revise and correct the named entity links found in a document. Such a tool can be used in cases where the annotations are critically important, such as in dealing with legal documents or parliamentary data[1] [9]. 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 [10], previously created by our research group.

4. Discussion

Using the tool presented in this paper in real time imposes limits on how much computing power and time can be used by the system. Some other significant challenges also apply to the approach and tool discussed. One of them is the use of Finnish language. One of our main objectives is to increase the digital presence of Finnish, making its processing tools more robust and full-featured. This, however, precludes the usage of LLMs which cannot handle this language satisfactorily. Another limitation is our focus on open-source tools, and previously discussed. Currently, the open-weights LLMs that produces best results when handling Finnish language are the Llama 3 models [11], which are the ones in use in this project.

[1]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.

Acknowledgements

Our work is part of the national FIN-CLARIAH research infrastructure programme, funded by the Research Council of Finland. This project has received funding from the European Union – NextGenerationEU instrument and is funded by the Research Council of Finland under grant number 346323. CSC – IT Center for Science has provided computational resources for our projects.

Biographies of Authors

Rafael Leal is a Doctoral Candidate in Computer Science at Aalto University. He is part of the Semantic Computing Research Group (SeCo) and his main research topic is the integration between NLP techniques and Linked Data.

Eero Hyvönen is professor of semantic media technology at the Aalto University, Department of Computer Science, and director of Helsinki Centre for Digital Humanities (HELDIG) at the University of Helsinki directing the Semantic Computing Research Group (SeCo) specializing on Semantic Web technologies and applications in Digital Humanities. He has published over 500 research articles and books and has got several international and national awards.

Annastiina Ahola is a Doctoral Researcher at the Department of Computer Science in Aalto University. She is a researcher in the Semantic Computing Research Group (SeCo) with her research focusing on developing tools and applications to aid in Digital Humanities research.

References

[1] J. L. Martinez-Rodriguez, A. Hogan, I. Lopez-Arevalo, Information extraction meets the semantic web: A survey, Semantic Web – Interoperability, Usability, Applicability 11 (2020) 255–335.

[2] E. Hyvönen, Digital humanities on the semantic web: Sampo model and portal series, Semantic Web – Interoperability, Usability, Applicability 14 (2022) 729–744. doi:10.3233/SW-190386.

[3] M. Tamper, R. Leal, L. Sinikallio, P. Leskinen, J. Tuominen, E. Hyvönen, Extracting knowledge from parliamentary debates for studying political culture and language, in: S. Tiwari, N. Mihindukulasooriya, F. Osborne, D. Kontokostas, J. D'Souza, M. Kejriwal (Eds.), 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), volume 3184, CEUR WS, 2022, pp. 70–79. URL: http://ceur-ws.org/Vol-3184/TEXT2KG_Paper_5.pdf, international Workshop on Knowledge Graph Generation from Text (TEXT2KG 2022).

[4] J. Luoma, M. Oinonen, M. Pyykönen, V. Laippala, S. Pyysalo, A Broad-coverage Corpus for Finnish Named Entity Recognition, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, 2020, pp. 4615–4624. URL: https://aclanthology.org/2020.lrec-1.567.

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive nlp tasks, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020.

[6] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[7] S. Chan, A. Santoro, A. Lampinen, J. Wang, A. Singh, P. Richemond, J. McClelland, F. Hill, Data distributional properties drive emergent in-context learning in transformers, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 18878–18891.

[8] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, volume 35, Curran Associates, Inc., 2022, pp. 22199–22213.

[9] E. Hyvönen, L. Sinikallio, P. Leskinen, S. Drobac, R. Leal, M. L. Mela, J. Tuominen, H. Poikkimäki, H. Rantala, Publishing and using parliamentary linked data on the Semantic Web: ParliamentSampo system for Parliament of Finland, 2024. In open review: https://www.semantic-web-journal.net/system/files/swj3605.pdf.

[10] A. Oksanen, M. Tamper, J. Tuominen, A. Hietanen, E. Hyvönen, Anoppi: A pseudonymization service for finnish court documents, in: M. Araszkiewicz, V. Rodríguez-Doncel (Eds.), Legal Knowledge and Information Systems. JURIX 2019: The Thirty-Second Annual Conference, IOS Press, 2019, pp. 251–254. doi:10.3233/FAIA190335.

[11] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. arXiv:2302.13971.