Extending the Finnish Linked Data Infrastructure with Natural Language Processing Services in FIN-CLARIAH

Tamper, Minna

2022-03-01

Extending the Finnish Linked Data Infrastructure with Natural Language Processing Services in FIN-CLARIAH

Minna Tamper[1,2], Jouni Tuominen[1,2,3] and Eero Hyvönen[1,2]

[1]*Semantic Computing Research Group (SeCo), Aalto University, Finland*
[2]*Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland*
[3]*Helsinki Institute for Social Sciences and Humanities (HSSH), University of Helsinki, Finland*

Abstract

The DARIAH-EU infrastructure for Digital Humanities (DH) is often focusing on using structured data for quantitative studies, while the EU-CLARIN infrastructure deals primarily with unstructured natural language texts. However, in DH research both texts and structured data are often needed. It therefore makes sense to develop and use both infrastructures together, as suggested in the Dutch CLARIAH programme and the corresponding FIN-CLARIAH initiative in Finland, a new part of the Finnish research infrastructure road map of the Academy of Finland. This poster paper introduces work in FIN-CLARIAH relating to the idea of integrating natural language processing (NLP) tools with the Linked Open Data (LOD) Infrastructure for Digital Humanities in Finland (LODI4DH). We present a plan for NLP services to be opened as part of the Linked Data Finland (LDF.fi) platform. The new services are used on the other hand for knowledge extraction from Finnish texts for weaving LOD, and on the other hand for language DH data analyses of the published datasets in applications in many domains, such as political culture. The extended LDF.fi platform will provide users with documented APIs for NLP services using unified output formats as well as software delivery as Docker containers, to lower the bar for deployment.

Keywords

knowledge extraction, natural language processing, linked data

1. Introduction

FIN-CLARIAH (2022–) is the premier Finnish digital research infrastructure development initiative for Social Sciences and Humanities (SSH) comprising two components:

1. FIN-CLARIN: Finnish dimension of the pan-European CLARIN[1] infrastructure
2. DARIAH-FI: Finnish dimension of the pan-European DARIAH[2] infrastructure

In their first common development project, the FIN-CLARIAH components seek to significantly broaden their mutual scope of digital SSH infrastructural support by consolidating and enhancing their resources with three major goals:

[1]https://www.clarin.eu
[2]https://www.dariah.eu

1. Reach beyond processing of spoken standard Finnish into colloquial speech
2. Cater to a broad range of SSH research needs for processing unstructured text
3. Facilitate research based on metadata

FIN-CLARIAH involves Finnish universities with research in SSH, including the coordinator University of Helsinki (Faculty of Arts, Faculty of Social Sciences, and National Library), CSC – IT Center for Science Ltd., Aalto University, Tampere University, University of Eastern Finland, University of Jyväskylä, and University of Turku. In addition, FIN-CLARIAH has as project collaborators University of Vaasa, University of Oulu, the Institute for the Languages of Finland, and the National Archives of Finland.

This poster paper introduces development work of the Semantic Computing Research Group (SeCo) in FIN-CLARIAH at the Aalto University and University of Helsinki, Helsinki Centre for Digital Humanities (HELDIG)[3], on integrating natural language processing (NLP) tools with the Linked Open Data (LOD) Infrastructure for Digital Humanities in Finland (LODI4DH)[4] [1]. In conclusion, related works are discussed and contributions of our work summarized.

2. Services for Weaving Linked Data from Texts

The NLP services to be opened to public use are targeted to knowledge extraction from Finnish texts, based on named entity recognition (NER), linking, and relation extraction [2]. The same NLP technology can also be used for pseudonymizing texts [3] needed due to the GDPR regulations[5] of EU.

The idea in our work is to reuse various existing NLP tools from other developers and re-purpose and re-package them for extracting Linked Data from unstructured texts. For example, for named entity recognition a model trained using FinBERT NER[6] [4] is used. The rule-based methods to be used include regular expressions (to find prescribed surface forms such as property codes and vehicle registration plates) and dictionaries, such as the Finnish person name ontology [5]. In addition, the Turku Neural Parser [6] is used to perform grammatical and morphological analysis on the whole document. NLP tools developed in our own earlier projects for the various Sampo portals [7] will also be re-used here.

The NLP tools will be used in FIN-CLARIAH for language analyses of texts and for DH analyses regarding their subject matter content. Application case study areas here include analysing the nearly million speeches 1907—2021 of the Parliament of Finland[7], biographical collections, such as the National Biography of Finland[8], and Finnish legislation and case law[9] published by the Ministry of Justice. Texts for these application areas are already available through the Sampo series of LOD services and portals[10] [7] but also other datasets from the participants of FIN-CLARIAH will be used on as-needed basis.

---

[3]https://seco.cs.aalto.fi/projects/fin-clariah/
[4]https://seco.cs.aalto.fi/projects/lodi4dh/
[5]https://europa.eu/youreurope/business/dealing-with-customers/data-protection/data-protection-gdpr/
[6]https://turkunlp.org/fin-ner.html
[7]https://seco.cs.aalto.fi/projects/semparl/
[8]https://seco.cs.aalto.fi/projects/biografiasampo/
[9]https://seco.cs.aalto.fi/projects/lakisampo/
[10]https://seco.cs.aalto.fi/applications/sampo/

The NLP services will be provided using the existing Linked Data Finland platform[11] that is in use in Finland for publishing Linked Data [8]. The forth-coming NLP services provide users with demo applications and web APIs, unified output formats (e.g., JSON, RDF), documentation, and software delivery as Docker container images, which lowers the bar for deployment. The containerized tools can be deployed as components in data processing workflows, and be scaled to support varying workloads. In addition to using and providing tools as containers, the portal enables testing the tools with custom input.

3. Discussion

Portals for NLP web services have been created before for many languages, e.g., the GATE Cloud [9]. In addition to web services, this kind of tools have also been packaged at into useful modules available for programming languages, such as Python, e.g., EstNLTK[12] [10], UralicNLP[13] [11], StanfordNLP[14] [12], and NLTK[15] [13]. However, the modules or packages for specific programming languages not only tie their users to the language but also require more technical skills and understanding. A benefit of service portals is that it can provide users with demo UIs and APIs through which the users can use the services more easily using the HTTP protocol and regardless of the programming language they use.

Methods for information extraction are surveyed in [2]. A novelty of our work is its focus on Finnish and on knowledge extraction from texts for linked data to be used in research for language analyses and studies in Digital Humanities. The tools and services planned to be included in the portal come from various NLP and DH projects that can be utilized to extract knowledge from different types of source texts. The services are distributed as containers and their results are converted into well-known output formats to ease deployment and improve usability in other DH projects.

Acknowledgments

References

[1] E. Hyvönen, Linked open data infrastructure for Digital Humanities in Finland, in: Proceedings of Digital Humanities in Nordic Countries (DHN 2020), CEUR-WS Proceedings, Vol. 2612, 2020, pp. 254–259. URL: http://ceur-ws.org/Vol-2612/short10.pdf.

---

[11] https://ldf.fi
[12] https://github.com/estnltk/estnltk
[13] https://github.com/mikahama/uralicNLP
[14] https://github.com/stanfordnlp/stanfordnlp
[15] https://github.com/nltk/nltk

[2] 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.

[3] A. Oksanen, M. Tamper, J. Tuominen, A. Hietanen, E. Hyvönen, Anoppi: A pseudonymization service for Finnish court documents, in: Legal Knowledge and Information Systems. JURIX 2019: The Thirty-second Annual Conference, IOS Press, 2019, pp. 251–254.

[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, Marseille, France, 2020, pp. 4615–4624. URL: https://aclanthology.org/2020.lrec-1.567.

[5] M. Tamper, P. Leskinen, J. Tuominen, E. Hyvönen, Modeling and publishing finnish person names as a linked open data ontology, in: 3rd Workshop on Humanities in the Semantic Web (WHiSe 2020), CEUR Workshop Proceedings, vol. 2695, 2020, pp. 3–14. URL: http://ceur-ws.org/Vol-2695/paper1.pdf.

[6] J. Kanerva, F. Ginter, N. Miekka, A. Leino, T. Salakoski, Turku neural parser pipeline: An end-to-end system for the CoNLL 2018 shared task, in: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, 2018, pp. 133–142.

[7] E. Hyvönen, Digital humanities on the Semantic Web: Sampo model and portal series, Semantic Web – Interoperability, Usability, Applicability (2021). URL: http://www.semantic-web-journal.net/content/digital-humanities-semantic-web-sampo-model-and-portal-series, submitted.

[8] E. Hyvönen, J. Tuominen, M. Alonen, E. Mäkelä, Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets, in: The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers, Springer, 2014, pp. 226–230. URL: https://doi.org/10.1007/978-3-319-11955-7_24.

[9] V. Tablan, I. Roberts, H. Cunningham, K. Bontcheva, Gatecloud. net: a platform for large-scale, open-source text processing on the cloud, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 371 (2013) 20120071.

[10] S. Laur, S. Orasmaa, D. Särg, P. Tammo, EstNLTK 1.6: Remastered Estonian NLP pipeline, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 7152–7160. URL: https://aclanthology.org/2020.lrec-1.884.

[11] M. Hämäläinen, UralicNLP: An NLP library for Uralic languages, Journal of Open Source Software 4 (2019) 1345. doi:10.21105/joss.01345.

[12] P. Qi, T. Dozat, Y. Zhang, C. D. Manning, Universal dependency parsing from scratch, in: Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 160–170. URL: https://nlp.stanford.edu/pubs/qi2018universal.pdf.

[13] S. Bird, E. Klein, E. Loper, Natural language processing with Python: analyzing text with the natural language toolkit, O'Reilly Media, Inc., 2009.