entity-fishing: a DARIAH entity recognition and disambiguation service

Foppiano, Luca
ALMAnaCH, Inria, Paris
luca.foppiano@inria.fr

Romary, Laurent
ALMAnaCH, Inria, Paris
laurent.romary@inria.fr

9 September 2018

Abstract

This paper presents an attempt to provide a generic named-entity recognition and disambiguation module (NERD) called *entity-fishing* as a stable online service that demonstrates the possible delivery of sustainable technical services within DARIAH, the European digital research infrastructure for the arts and humanities. Deployed as part of the national infrastructure Huma-Num in France, this service provides an efficient state-of-the-art implementation coupled with standardised interfaces allowing an easy deployment on a variety of potential digital humanities contexts. The topics of accessibility and sustainability have been long discussed in the attempt of providing some best practices in the widely fragmented ecosystem of the DARIAH research infrastructure. The history of *entity-fishing* has been mentioned as an example of good practice: initially developed in the context of the FP9 CENDARI,[1] the project was well received by the user community and continued to be further developed within the H2020 HIRMEOS project where several open access publishers have integrated the service to their collections of published monographs as a means to enhance retrieval and access.

entity-fishing implements entity extraction as well as disambiguation against Wikipedia and Wikidata entries. The service is accessible through a REST API which allows easier and seamless integration,

1. Patrice Lopez, Alexander Meyer, and Laurent Romary, *CENDARI Virtual Research Environment & Named Entity Recognition techniques*, Grenzen überschreiten – Digitale Geisteswissenschaft heute und morgen, Poster, Einstein-Zirkel Digital Humanities, February 2014, https://hal.inria.fr/hal-01577975.

language independent and stable convention and a widely used service oriented architecture (SOA) design. Input and output data are carried out over a query data model with a defined structure providing flexibility to support the processing of partially annotated text or the repartition of text over several queries. The interface implements a variety of functionalities, like language recognition,[2] sentence segmentation and modules for accessing and looking up concepts in the knowledge base. The API itself integrates more advanced contextual parametrisation or ranked outputs, allowing for the resilient integration in various possible use cases. The *entity-fishing* API has been used as a concrete use case[3] to draft the experimental stand-off proposal, which has been submitted for integration into the TEI guidelines.[4] The representation is also compliant with the Web Annotation Data Model[5] (WADM).

In this paper we aim at describing the functionalities of the service as a reference contribution to the subject of web-based NERD services. In order to cover all aspects, the architecture is structured to provide two complementary viewpoints. First, we discuss the system from the data angle, detailing the workflow from input to output and unpacking each building box in the processing flow. Secondly, with a more academic approach, we provide a transversal schema of the different components taking into account non-functional requirements in order to facilitate the discovery of bottlenecks, hotspots and weaknesses. The attempt here is to give a description of the tool and, at the same time, a technical software engineering analysis which will help the reader to understand our choice for the resources allocated in the infrastructure.

Thanks to the work of million of volunteers, Wikipedia has reached today stability and completeness that leave no usable alternatives on the market (considering also the licence aspect). The launch of Wikidata in 2010 have completed the picture with a complementary language independent meta-model which is becoming the scientific reference for many disciplines[6]. After providing an introduction to Wikipedia and Wikidata, we describe the knowledge base: the data organisation, the *entity-fishing* process to exploit it and the way it is

2. Shuyo Nakatani, *Language Detection Library for Java*, 2010, `https://github.com/shuyo/language-detection`.

3. https://hedgehog-web.herokuapp.com/json2xml/index.html

4. Piotr Banski et al., *Wake up, standOff!*, TEI Conference 2016, September 2016, `https://hal.inria.fr/hal-01374102`.

5. https://www.w3.org/TR/annotation-model/

6. https://blog.wikimedia.de/2014/10/22/establishing-wikidata-as-the-central-hub-for-linked-open-life-science-data/

built from nightly dumps using an offline process.

We conclude the paper by presenting our solution for the service deployment: how and which the resources where allocated. The service has been in production since Q3 of 2017, and extensively used by the H2020 HIRMEOS partners during the integration with the publishing platforms. We believe we have strived to provide the best performances with the minimum amount of resources. Thanks to the Huma-num infrastructure we still have the possibility to scale up the infrastructure as needed, for example to support an increase of demand or temporary needs to process huge backlog of documents. On the long term, thanks to this sustainable environment, we are planning to keep delivering the service far beyond the end of the H2020 HIRMEOS project.

References

Banski, Piotr, Bertrand Gaiffe, Patrice Lopez, Simon Meoni, Laurent Romary, Thomas Schmidt, Peter Stadler, and Andreas Witt. *Wake up, standOff!* TEI Conference 2016, September 2016. `https://hal.inria.fr/hal-01374102`.

Lopez, Patrice, Alexander Meyer, and Laurent Romary. *CENDARI Virtual Research Environment & Named Entity Recognition techniques*. Grenzen überschreiten – Digitale Geisteswissenschaft heute und morgen. Poster. Einstein-Zirkel Digital Humanities, February 2014. `https://hal.inria.fr/hal-01577975`.

Nakatani, Shuyo. *Language Detection Library for Java*, 2010. `https://github.com/shuyo/language-detection`.