Manual Validation of Automatic Annotation build by AlvisAE: Plant Health Bulletin Use Case

Marine COURTIN, Stephan BERNARD, Robert BOSSY, Catherine ROUSSEY

The ANR project "Data to Knowledge in Agronomy and Biodiversity" (D2KAB) (www.d2kab.org) aims to create a framework to turn agronomy and biodiversity data into knowledge –semantically described, interoperable, actionable, open– and investigate scientific methods and tools to exploit this knowledge for applications in science & agriculture. We shall provide the means –ontologies, linked open data resources and knowledge graphs– for agronomy and agriculture domains to embrace the semantic Web by exploiting and producing FAIR data. One of the tasks of the project focuses on the analysis of a particular corpus of French agricultural alert bulletins: the official Plant Health Bulletins (PHB – *bulletins de santé du végétal*, in French).

An archive of PHBs related to grapevine published as linked open data is constantly updated. It contains original PDF files and their HTML versions, plus a SPARQL endpoint available for querying the data. We apply Natural Language Processing techniques on these HTML documents to produce a set of PHB's annotations describing plot observations related to growth stages and pest attacks. An RDF annotation links HTML text segments with elements of semantic resources, all hosted in AgroPortal:

- French Crop Usage (FCU) thesaurus developed by INRAE-TSCF,
- BBCH-based Plant Phenological Description Ontology and associated knowledge graphs developed by D2KAB consortium and associates (Elzeard, IFV),
- TAXREF-LD developed by CNRS-I3S for the MNHN,
- a new disease resource about grapevine developed by D2KAB consortium and associates.

PHB documents were automatically annotated using unsupervised methods that leverage existing lexicons and the consortium expertise on grapevine and Natural Language Processing. The annotations will be manually validated by experts (4 INRAE agronomists) in order to provide high quality annotations. The whole pipeline uses tools developed at MaIAGE (INRAE). AlvisNLP [Ba and Bossy 2016] is a modular and customizable automatic corpus processing engine that integrates a library that encompasses all aspects of NLP (tokenization, pattern matching, lexicon search, machine learning) . AlvisAE [Papazian and al, 2012] is a Web application dedicated to manual annotation and validation that supports the annotation of entities, relations, coreferences, simple or double annotations.

We will present with few examples the evaluation of automatic annotation. In the future, the annotated corpus will be queried in order to express the evolution of crops and pest attacks in different french areas.

References

Mouhamadou Ba, Robert Bossy. Interoperability of corpus processing workflow engines: the case of. AlvisNLP/ML in OpenMinTeD. *Meeting of working Group Medicago sativa*, May 2016, Portoroz, Slovenia. ⟨hal-01455853⟩

Frederic Papazian, Robert R. Bossy, Claire Nédellec. AlvisAE: a collaborative Web text annotation editor for knowledge acquisition. *LAW VI '12 - Sixth Linguistic Annotation Workshop*, Jul 2012, Jeju, South Korea. pp.184. ⟨hal-02748212⟩