Information extraction in plant biology

Claire Nédellec, Robert Bossy
INRAE MaIAGE, Université Paris-Saclay, Jouy-en-Josas, France
{firstname.lastname}@inrae.fr

Article Type: *Annotated datasets - training and evaluating LLMs*

Summary

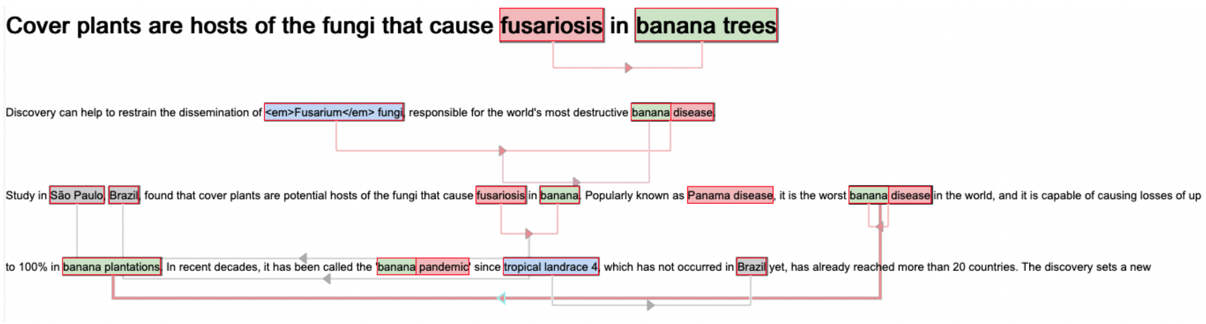
We intend to finalize the preparation of the *Plant Health Surveillance* task of the EPOP (*Epidemiomonitoring Of Plant*) corpus and the *Wheat breeding* task of the Taec corpus (*The Triticum aestivum trait Corpus*) with a view to its dissemination in two complementary stages: (1) to convert the corpora annotations into a standard format and expose the data on the PubAnnotation portal, (2) to carry out corpus relevance tests with international participants of BLAH8 using state-of-the-art methods.

Description of the project

The Triticum aestivum trait Corpus **(Taec)** and *Epidemiomonitoring Of Plant* (**EPOP)** are two datasets in the field of plant biology that have been manually annotated by INRAE experts using a sound methodology (double-blind annotation, detailed annotation guidelines). Information extraction for plant biology is an area poorly endowed with annotated corpora. For *Arabidopsis thaliana*, there have been some recent Information Extraction initiatives, such as the *KnownLeaf* literature curation system (Szakonyi et al., 2015) and the *SeeDev* reference corpus (Chaix et al., 2016). The *Knownleaf* corpus focuses on the regulatory mechanisms of leaf growth and development, and key genes related to relevant mutant phenotypes The *SeeDev* corpus concentrates on seed development described at the molecular level. The Taec and EPOP corpora focus on a different topic which is crop plant traits and health.

We also introduce the **IEval** tool to evaluate predictions made by automatic systems on these datasets (https://github.com/Bibliome/IEval). The corpora are built with the purpose of training and evaluating Natural Language Processing (NLP) methods. This proposal aims to integrate these two datasets into the PubAnnotation repository. We plan to assess the relevance of the annotated corpora with BLAH8 participants by state-of-the-art methods with a view to preparing a joint publication on the dataset and machine learning tests.

The theme of the **EPOP corpus** is the evolution of knowledge in crop plant health through international news translated into English. The NLP task consists of extracting mentions of observations represented by biological interaction events between monitored pests, the diseases they cause, and their biological interactions with their vectors and host plants in a given place and at a given date. The aim is twofold: to support health and scientific monitoring using automatic methods, and to provide the BioNLP community with a new reference corpus comprising biological and spatio-temporal relationships in the field of plants.To this end, we have prepared the manually annotated EPOP corpus. The annotation project has mainly involved the MaIAGE laboratory, the National Plant Health Platform, and more than 30 experts in plant health. The EPOP corpus contains 264 news in which text entities and their relationships are semantically annotated using domain references. The NLP task involves the recognition and normalization of entities and the extraction of relationships between these entities. The entities are normalized by semantic resources (i.e. OntoBiotope ontology, Geonames repository and NCBI taxonomy), whose large size is an original feature of this task.
The annotation schema comprises 8 entities, 8 oriented binary relations, and the n-ary relations composed of these binary relations (see the example below).



Cover plants are hosts of the fungi that cause fusariosis in banana trees

Discovery can help to restrain the dissemination of <em>Fusarium</em> fungi, responsible for the world's most destructive banana disease.

Study in São Paulo, Brazil, found that cover plants are potential hosts of the fungi that cause fusariosis in banana. Popularly known as Panama disease, it is the worst banana disease in the world, and it is capable of causing losses of up to 100% in banana plantations. In recent decades, it has been called the 'banana pandemic' since tropical landrace 4, which has not occurred in Brazil yet, has already reached more than 20 countries. The discovery sets a new

The **Taec dataset** is a new gold standard for traits and phenotypes of wheat. Wheat varieties show a large diversity of traits and phenotypes. Linking them to genetic variability is essential for shorter and more efficient wheat breeding programs. Newly desirable wheat variety traits include disease resistance to reduce pesticide use, adaptation to climate change, resistance to heat and drought stresses, or low gluten content of grains. Wheat breeding experiments are documented by a large body of scientific literature and observational data obtained in-field and under controlled conditions. The cross-referencing of complementary information from the literature and observational data is essential to the study of the genotype-phenotype relationship and to the improvement of wheat selection.
The scientific literature describes much information about the genotype-phenotype relationship. However, the variety of expressions used to refer to traits and phenotype values in scientific articles is a hindrance to finding information and cross-referencing it. The *Triticum aestivum trait Corpus* is a new gold standard for training and evaluating named entity recognition and entity-linking methods in plant phenotype literature. It consists of 540 PubMed references fully annotated for trait, phenotype, and species named entities using the Wheat Trait and Phenotype Ontology (Nedellec et al., 2020) and the species taxonomy of the National Center for Biotechnology Information.

The Taec datasets is available at
https://github.com/Bibliome/Integrating-plant-biology-datasets-and-evaluation-tool-into-PubAnnotation
in BioNLP-ST format. The EPOP corpus will be converted in BioNLP-ST format and made available in the Github projet by the time of the workshop. We plan on developing a format converter for the integration into PubAnnotation for both datasets. The evaluation tool that we plan to develop will allow users to evaluate their predictions on the subtasks of EPOP and Taec. Evaluation metrics include the common precision, recall, and F1 scores, but also slot error rate (SER) and similarity-based measures which can be used to evaluate entity recognition and normalization. An Information Extraction evaluation tool will be configured, i.e. parametrized to compute relevant metrics for the two data sets. It also features a wide range of filters and similarity functions that allow it to compute more detailed evaluations of predictions, such as fuzzy match of named entity prediction or normalization. The evaluation tool has a command-line interface, a Web UI, and a REST API. The services for the evaluation of the predictions of two datasets will be documented in the GitHub repository.

Motivation

PubAnnotation is a repository to share annotation datasets and tools that help foster annotation efforts in the biomedical domain. We would like to add to this shared effort by making available our benchmark datasets and evaluation tool. The PubAnnotation repository already contains a number of datasets among which those proposed within the BioNLP-ST and BioNLP-OST framework, and adding EPOP and Taec would complement this collection. We aim to attract new users and promote the development of information extraction systems for plant biology domains.
We intend to finalize the preparation of the PHS (*Plant Health Surveillance*) task of the EPOP corpus and the *Wheat breeding* (WB) task of the Taec corpus with a view to its dissemination and submission as shared tasks (e.g. at CLEF (*Conference and Labs of the Evaluation Forum*)), in two complementary stages: (1) to convert the corpora annotations into a standard format and expose the data on the PubAnnotation portal, (2) to carry out corpus relevance tests with international participants of BLAH8 using state-of-the-art methods.

Specific Goals

BioNLP-ST to PubAnnotation format converter. The BioNLP-ST format is quite common in the BioNLP community, there are several datasets available, and the format is very close to that of the widely used annotation visualization tool Brat. We plan to make this converter available in an open license.

- Anticipated challenges: URI design, discontinuous entities, entity normalization representation, n-ary relations.

Upload the datasets to PubAnnotation.

- Anticipated challenges: conversion mistakes.

Configure the evaluation tool. Adaptation of the evaluation tool to the peculiarities of both tasks.

Dataset split. The two corpora will be divided into three parts: training, development and testing ensuring a uniform distribution of the annotation types. The test subset will remain unpublished to avoid overtraining.

Involvement of BLAH8 participants in the application of NER, NEL, and RE methods. Sharing annotation guidelines and characteristics of the task challenges. Possibly revising dataset annotations according to methods prediction error analysis.

Hackathon Environment

The BLAH Hackathon is a great venue to work on the integration of our datasets and evaluation tool. In a hackathon framework, we are able to dedicate ourselves entirely to a specific task. We can also benefit from the expertise of other BLAH participants, especially on PubAnnotation formats and protocols.
BLAH Hackathon will also be a great place to promote our new corpora and identify potential flaws before publication and shared task organization at CLEF.

Evaluation

The dataset conversion and registration will be evaluated by exploiting the PubAnnotation native visualization tool. Since we have a deep insight into the datasets, we will be able to check that the converted annotations are faithful to the official dataset. In particular, we will target “difficult cases” registered during the annotation process.

The in-depth analysis of LLM prediction errors will contribute to improving the quality of the two datasets.

Travel Support

We do not require travel support to fit the BLAH8 financial constraints. Arnaud Ferré from our Bibliome group is applying for full travel support.

References

Chaix E, Dubreucq B, Fatihi A, et al. Overview of the Regulatory Network of Plant Seed Development (SeeDev) Task at the BioNLP Shared Task 2016. In: *Proceedings of the 4th BioNLP Shared Task Workshop*. Association for Computational Linguistics; 2016:1-11. doi:10.18653/v1/W16-3001

Nédellec C, Ibanescu L, Bossy R, Sourdille P. WTO, an ontology for wheat traits and phenotypes in scientific publications. *Genomics Inform*. 2020;18:e14. doi:10.5808/GI.2020.18.2.e14

C. Nédellec, L.Deleger, C. Sauvion. Guidelines for the Annotation of the *Taec* Corpus - Phenotype and Trait information in Wheat. INRAE MaIAGE. 2023. ⟨hal-04118664⟩ https://doi.org/10.57745/GCYG3Q

Nédellec, Claire; Sauvion, Clara; Deléger, Louise; Bossy, Robert; Zweigenbaum, Léonard, 2023, "Triticum aestivum trait Corpus", https://doi.org/10.57745/GCYG3Q, Recherche Data Gouv, V1

Dóra Szakonyi, Sofie Van Landeghem, Katja Baerenfaller, Lieven Baeyens, Jonas Blomme, et al.. The KnownLeaf literature curation system captures knowledge about Arabidopsis leaf growth and development and facilitates

integrated data mining, *Current Plant Biology*, Volume 2, 2015, Pages 1-11, ISSN 2214-6628, https://doi.org/10.1016/j.cpb.2014.12.002.