Extracting Crop Model Parameters from Literature Using Natural Language Processing

Maria Alexeeva, Vijaya Raj Joshi, Hubert Kanyamahanga, Isaac Kobby Anni, Keith Alcock, Gerrit Hoogenboom, Mihai Surdeanu

Decision Support System for Agrotechnology Transfer (DSSAT) is a software application program to simulate crop growth, development, and yield. DSSAT consists of crop models for more than 40 different crops. For DSSAT application, it requires a lot of input data on crop cultivar, crop management practices, weather, and soil properties. However, when applying models to new locations, these input parameters are difficult and expensive to obtain. In our work, we overcome this limitation by automatically extracting parameters from scientific publications and reports related to agriculture in the regions of interest. At the core of our approach is a machine reading system built using the rule-based information extraction framework named Odin, used in combination with preprocessing (pdf to text conversion and text preprocessing) and post-processing (redundancy filtering, binarization, etc) components. The intuition behind our method is that we extract references of variables (or parameters of crop prediction systems) such as “sowing” and we link them to the corresponding values mentioned in text such as actual planting dates. We extract several types of information, including crop varieties, yield amounts, area sizes, and more. For every event that we extract, we provide available context, e.g., location, season, date, crop, fertilizer, etc. This allows for better filtering of extractions in downstream tasks. For instance, for planting date extractions, knowing the associated crop can help the user select planting events for specific types of crops that they need to model. We evaluate the quality of extractions (excluding the quality of context assignment) on two sets of papers for the same region (Senegal) on two types of crops: rice (6 papers) and peanut (12 papers). The accuracy—the proportion of extractions that we judge to be correct out of all extractions—is 0.84 for rice based on 334 extractions and 0.95 for peanuts based on 224 extractions.