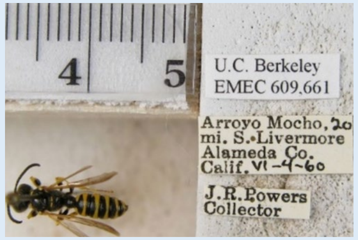
HuMaIN

Human- and Machine-Intelligent Network of Software Elements for Cost-Effective Scientific Data Digitization

Ícaro Alzuru, Andréa Matsunaga, Maurício Tsugawa and José A.B. Fortes

Problems and Research Questions



Darwin Core terms:
Collected By: A. A. Heller, P. B. Kennedy
State: California
County: Plumas
Identified By: James L. Reveal

121,428,342
Specimen Records
31,871,863
Media Records
1,621
Recordsets

Why digitization matters
More about what we do and why

- Problem: **Efficient** Information Extraction (IE) from biocollections.
- How do **crowdsourcing interfaces** affect output quality & crowd sentiment?
- How can **quality** of automated IE match quality of IE by humans?
- How to create **IE workflows** that combine human and machine tasks?
- Can **general approaches** for IE and confidence estimation be pursued?

Research Findings

Human-like quality but faster…

SELFIE	Avg. change
Quality	-0.3%
Duration	-27.2%
Crowd--sourcing Reduction	-32.2%

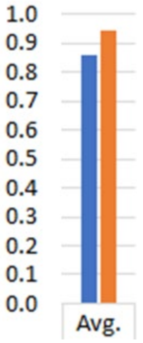
cheaper…

OCR Ensemble for Text Extraction versus	Total savings
Dynamic Human-Machine Consensus	73.35%
Hybrid Transcriber /Reviewer	78.78%

trained w/ "good" IE data…

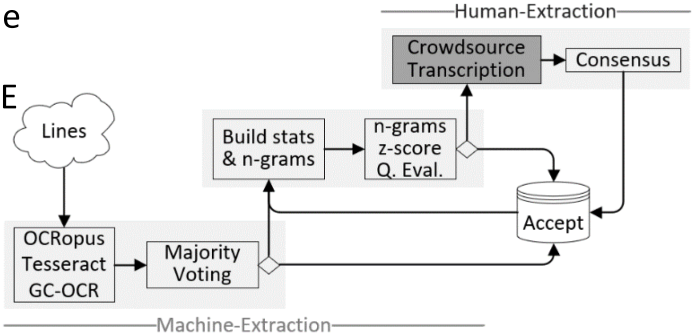
Recall: 0.859

Similarity: 0.941



Approaches and Methods

- **SELFIE**: Self-aware IE
- **Ensemble** of OCRs engines for the estimation of confidence in IE.
- Use of available **IE data to train** IE and confidence estimation methods.
- **Human-in-the-loop** methods: iterative training and improvement of IE quality and confidence estimation.



Conclusions and Deliverables

- Human-machine workflows for IE of DC terms from specimens' images.
  http://humain.acis.ufl.edu **https://github.com/acislab/HuMaIN**
- Used with biocollections from iDigBio, University of Australia, and WeDigBio.
- Ensembles of OCR, Human-in-the-loop, Named-entity Recognition, and Frequency Lists successfully tested for IE and IE confidence estimation.
- HuMaiIN data/methods can be tried/extended with open-source simulator:
  https://github.com/acislab/HuMaIN_Simulator

NSF SI^2 PI Workshop Seattle, WA – February 13-14, 2020