HuMaIN: Human- and Machine-Intelligent Network of Software Elements

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

ACIS ADVANCED COMPUTING & INFORMATION SYSTEMS

Motivation & Approach

Despite the advances on Natural Language Processing (NLP) and Machine Learning, the digitization of scientific data is usually performed using **crowdsourcing** because of the low confidence on the results generated by automated approaches.

HuMaIN's goal is to improve scientific-digitization efficiency through the **integration and synergistic cooperation of human and machine processes**. In HuMaIN:

- **Self-aware Information Extraction (IE) tasks** partially substitute crowdsourcing. They are able to identify when human help is really needed.
- Crowdsourcing results (from human work) are used to train & improve machine tasks
- Due to the importance of human participation, best practices in the design of crowdsourcing tasks have been identified, considering the quality of the results, the time required to extract the information, and the crowd sentiment.

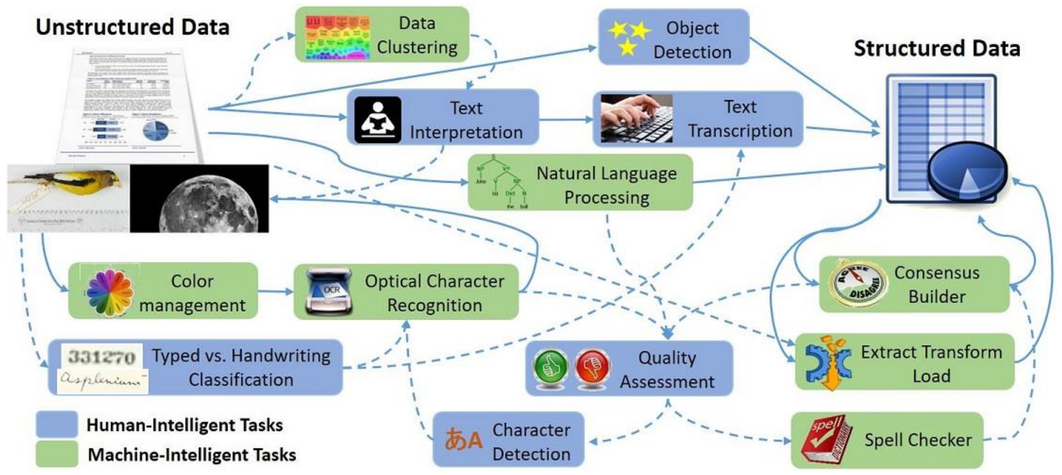

Figure 3. Hybrid (Human-Machine) IE of Scientific Data

Progress and Results

- First years: our results showed that hybrid (human-machine) IE approaches can exploit the benefits of both alternatives to create more efficient IE projects.
- Last year: we developed and tested a Self-aware IE (SELFIE) Model for efficient hybrid (human-machine) digitization of biocollections' labels.
- A study was conducted on how task granularity and Data Entry Methods (DEMs) affect the results' quality, duration, and crowd sentiment on IE tasks.
- We implemented OCROpus and Tesseract web services to enable massive and distributed IE. Code available at: https://github.com/acislab/HuMaIN_Microservices
- Paper presented at the 2017 IEEE 13th International Conference on eScience: **SELFIE: Self-aware Information Extraction from Digitized Biocollections**. Code and raw results at https://github.com/acislab/HuMaIN_Self-aware_Information_Extraction
- Paper presented at the 3rd IEEE Collaboration and Internet Computing Conference: **Task Design and Crowd Sentiment in Biocollections Information Extraction.**

Self-aware Information Extraction (SELFIE) from Biocollections

Components of SELFIE:

- SELFIE Workflow
- Self-aware Processes (SaPs)
- Self-aware Tasks (SaTs)
- Tasks

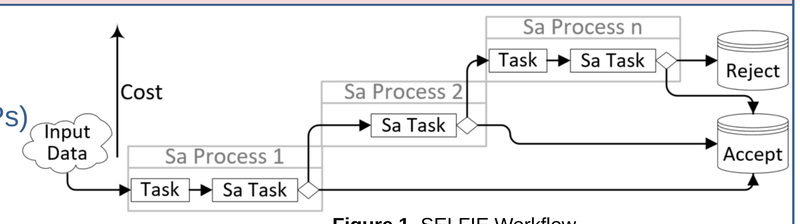

Figure 1. SELFIE Workflow

For each term to extract, a **SELFIE workflow** is defined: a sequence of IE steps with the capacity to **self-evaluate** and **adapt** to ensure the fulfillment of the IE goals.

SaPs are human or machine **IE alternatives**, organized in **incremental-cost order** to set the SELFIE workflow. The cost is a function of performance variables defined by the workflow designer.

SaTs extract and **evaluate** candidate values, taking the most appropriate **action**: accepting the best candidate or sending the image to be processed to the next SaP.

Part	Input	Adaptable Script/program	Adaptable Acceptance Method	Outputs
Example	*Image x*	*/path/script1.py*	*[0,b) -> Task y* *[b,1] -> Accept*	*Image x* *Value, Confidence*

Figure 2. Parts of a Self-aware Task

Tasks are other data manipulation jobs required for the SaPs.

The crowdsourcing data was obtained using ad-hoc interfaces during on-site crowdsourcing sessions (IRB 201600517). http://humain.acis.ufl.edu/aware/

SELFIE Experiments:

- Event date: Alphanumeric with defined patterns. NLP: Regular expressions.
- Scientific name: Textual known field. NLP: Suffixes (patterns), Sequential search.
- Recorded by: Textual unknown field. NLP: Dynamically created dictionary + search.

Table I. Results' improvement obtained by SELFIE, when compared to the Human-only IE approach.

	Event date	Scientific name	Recorded by	Avg.
Quality Improvement	-1.9%	1.6%	-0.6%	**-0.3%**
Duration Reduction	45.8%	15.3%	20.4%	**27.2%**
Crowdsourcing Reduction	48.0%	25.0%	23.5%	**32.2%**

Work in Progress

- Self-aware Information Extraction:
  - Improved accuracy (smarter) acceptance criteria for Self-aware Tasks.
  - Predictive probabilistic model.
- Alternative DEMs are being evaluated for the design of friendlier and more efficient crowdsourcing interfaces.

SELFIE Example – *Event date* Extraction

The *Event date* is the date when the specimen was collected.

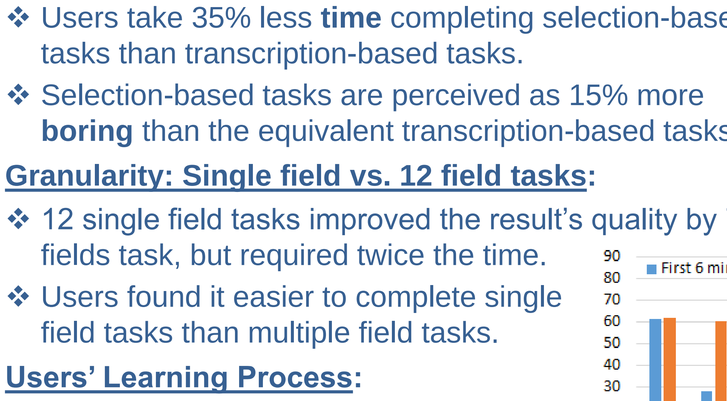

Figure 4. SELFIE workflow for *Event date*.

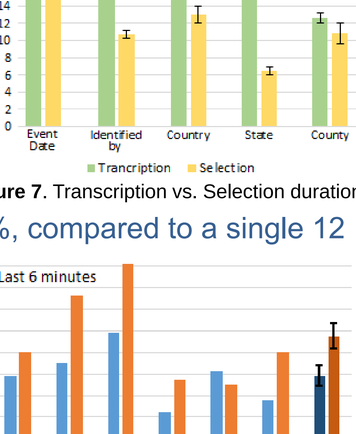

Figure 5. EMEC 609,661.

OCR: Tesseract generates a text file with presumably all the text in the image.

Reg. Expr.: A regular-expression-based script analyzes OCR text output and returns the earliest date found. The self-aware component accepts/rejects candidate values. If no value is accepted, the *Event date* of the image is sent to be crowdsourced.

Crowdsource Transcription: Volunteers transcribe the *Event date* of every image.

Consensus: Decides the final value for *Event date* based on the transcriptions.

Table II. Similarity to experts' transcription (Quality):

SaP/SELFIE	# Accepted	Similarity	SEM	Std. Dev.
Machine-only	48	0.934	0.024	0.167
Human-only	51	0.971	0.022	0.155
SELFIE	99	0.953	0.016	0.162

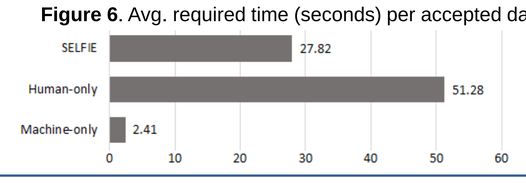

Figure 6. Avg. required time (seconds) per accepted date.

Task Design and Crowd Sentiment in IE from Biocollections

Data Entry Methods: Transcription vs. Selection:

- Selection-based tasks generate results of 7.7% higher **quality** than transcription-based tasks.
- Users take 35% less **time** completing selection-based tasks than transcription-based tasks.
- Selection-based tasks are perceived as 15% more **boring** than the equivalent transcription-based tasks.

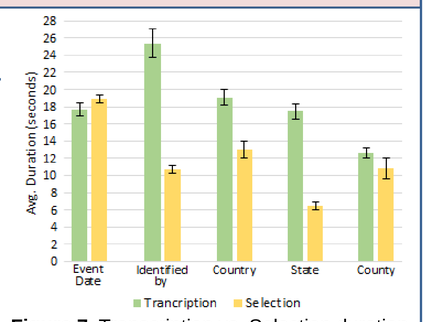

Figure 7. Transcription vs. Selection duration

Granularity: Single field vs. 12 field tasks:

- 12 single field tasks improved the result's quality by 7.25%, compared to a single 12 fields task, but required twice the time.
- Users found it easier to complete single field tasks than multiple field tasks.

Users' Learning Process:

When the first and the last 6 minutes of crowdsourcing are compared, users extract:

- Metadata of an equivalent quality.
- 1.5x more values in the last 6 minutes.

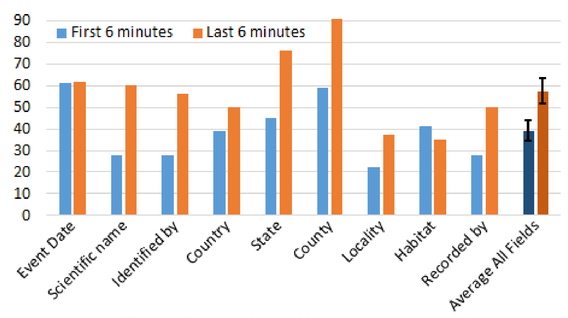

Figure 8. Number of images processed

Summary and Conclusions

- The SELFIE model has been devised for the efficient integration of human and machine IE processes.
- SELFIE has been shown to significantly reduce the number of crowdsourcing sessions required and the duration of the IE projects, while generating results of a quality equivalent to the generated by the human-only approach.
- A study of how task design affects crowdsourcing results was conducted:
  - Selection returns faster and higher quality results than Transcription, but is less fun.
  - Single-field tasks generate better quality results than many-field tasks.

NSF SI2 PI Workshop : April 30 – May 1, 2018

This material is based in part upon work supported by the National Science Foundation under Grant No. 1535086. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

UF UNIVERSITY of FLORIDA