# Document Analysis Algorithm Contributions in End-to-End Applications: Report on the ICDAR 2011 Contest

Bart Lamiroy
*LORIA*
*Nancy Université – INPL*
*Nancy, France*
*Bart.Lamiroy@loria.fr*

Daniel Lopresti
*Computer Science and Engineering*
*Lehigh University*
*Bethlehem, PA 18015, USA*
*lopresti@cse.lehigh.edu*

Tao Sun
*Computer Science and Engineering*
*Lehigh University*
*Bethlehem, PA 18015, USA*

***Abstract*—This contest aims to provide a metric giving indications on the influence of individual document analysis stages to overall end-to-end applications. Contestants are provided with a full, working pipeline which operates on a page image to extract useful information. The pipeline is built with clearly identified analysis stages (e.g. binarization, skew detection, layout analysis, OCR ...) that have a formalized input and output. Contestants are invited to contribute their own algorithms as an alternative to one or more of the initially provided stages. The evaluation measures the overall impact of the contributed algorithm on the final (end-of-pipeline) output.**

***Keywords*-benchmark; web services; document analysis; performance evaluation;**

## I. Introduction

This is the first edition of a contest that tries to measure impact of individual, focused research contributions to more broad document analysis problems. In order to achieve this, we build upon two main concepts: the ability to express full scale document analysis workflows in a flexible and extensible way [1] and the access to a data repository capable of tracking usage of the data and storing the resulting provenance so as to certify the results of the evaluation. [2].

The expected impacts of this contest are:

1) It will allow the community, over time, to identify where significant contributions are taking place, where bottlenecks are occurring and what hard problems are still unsolved (significance bearing the meaning of "end-user application significant") . These measures will not be based on publication metrics and activity, but on tangible results of running algorithms. Contrary to usual contests, where contestants are trying to provide a solution to a whole process, here they can focus on one part of the process only, the overall evaluation being done for all combinations over all contributions.
2) Given the specific setup that is used (http://dae.cse.lehigh.edu), users will be able to build upon previous work and published state-of-the art algorithms with a known track record, thus, in time, creating greater re-usability, and promoting objective evaluation of published methods. This is due to the fact that the contest platform is built on flexible web-service enabled storage infrastructure that can host and interact with a wide variety of execution formats [1]. Since, to compete, candidates need only to conform to specific input-output patterns, the entry level cost to interoperability, re-usability and sustainability is much lower.
3) The supporting infrastructure of the contest guarantees that contributions can be kept available over time and replayed on new datasets. This permits the effective monitoring of long term trends on the one hand, as well as extending the footprint of the contest to larger, more comprehensive end-to-end applications, always maintaining the knowledge of previous contributions.

## II. 2011 Specifics

Some of our intended goals for this contest can only be achieved over time. For the first edition, we have chosen to focus on the following points: Since this is the first time the contest is run, it convenes to scale it progressively, carefully assessing the feasibility of all goals. For this first edition, we have specifically focused on the following points:

1) The contest requires participants to expose their contributions as a web service (either hosted by themselves, either hosted by the contest organizers). This architecture is a testbed to evaluate whether a web service architecture may play a useful role in advancing document analysis research, especially by establishing how researchers will use such a system. Furthermore, while it is clearly an easy, flexible and ubiquitous way to create full pipeline prototypes, there is no doubt that custom developed and integrated applications will be more time/memory efficient. This contest will also be an opportunity to study the scalability in this domain.
2) Feedback from participants will provide useful input on the impact of this approach, beyond the vision we

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.302

have outlined earlier, and highlight its benefits and its limitations.

3) The need to bootstrap an initial set of algorithms for the pipeline, as well as a deliberate attempt to inspire contributors to improve the available stages, makes the current contest environment somewhat simplistic. Indeed, the proposed pipeline (Binarization – Text Layout Segmentation – OCR – Named Entity Detection) is rather simplistic and may prove limiting in some ways. Its construction is a bet on the current maturity of state-of-the art DIA methods, and shall be extended in function of the received contributions and their results. As stated before, there is currently no real cartographic overview of how advances in the document analysis domain contribute to the quality of end-to-end applications.

## III. Contest Description

For the contest, participants have been provided with a a full operational five-stage workflow, available from http://dae.cse.lehigh.edu/DAE/sites/default/files/ICDARContest.t2flow. In order to execute the workflow, participants were invited to use Taverna [3]. The individual stages of the workflow were hosted web-services on http://dae.cse.lehigh.edu/DAE/services/soap and consisted of:

Document Selection:
: Random document images are extracted from the contest data repository and provided as input to the next stage. No further assumption on image format or encoding is given.

Document Binarization:
: Document images from the previous stage are treated to produce a binarized version as output. The output file can be in any open and commonly used bitonal format (.tif, pbm, ...)

Layout Analysis/Text Localization:
: The bitonal images from the previous step are output either as a single document image, stripped from pictures, illustrations and, generally speaking, anything that is not text, or are output as a set of images, corresponding to the text blocs of the original document image.

OCR:
: The previously produced documents images are transcribed into text. The output is a plain text file without any formating characters other than whitespace.

Named Entity Detection:
: The text files are parsed and named entities such as places, people and organizations are tagged as such.

Participants were invided to provide access to implementations of one (or more) of the previous stages. Multiple entries for the same or for different categories were allowed. In order to test their contributions, participants were provided with standard, of-the-shelf, versions of each stage. Binarization was provided through the ImageMagick convert[1] program, a zero-effect segmentation web-service was provided (*i.e.* input images are simply echoed, without any segmentation), two OCR engines were made available: Ocrad[2] and Tesseract 2.04 [4][3], and the named entity detection software was the Stanford Named Entity Recognizer [5].

## IV. Contributions

Two research groups participated in the contest:

- the French EPITA Research and Development Laboratory[4],
- the Amercian NCI/CADD group, a research unit within the Chemical Biology Laboratory, part of the Molecular Discovery Program at the National Cancer Institute[5].

EPITA contributed three algorithm implementations, all of which were hosted as web-services at their own facilities[6]:

- an implementation of the Sauvola binarization algorithm [7] (EPITA_BI),
- a "generic" document text segmentation algorithm (EPITA_SEG),
- a specific document text segmentation algorithm that also participated in the ICDAR 2011 Historical Document Layout Analysis Contest (EPITA_HDOC).

NCI/CADD contributed two algorithm implementations, both of which were hosted as web-services by the contest organizers in virtualized environments [8]:

- a binarization algorithm (NCI_BI),
- a segmentation algorithm (NCI_SEG).

These five contributions, combined with the provided default algorithms therefore offer three binarization choices, four segmentation options and two OCRs, resulting in 24 different ways of extracting the named entities. Figure 1 depicts the full Taverna workflow[6] integrating all possible combinations that was used for running the contest.

## V. Contest Results and Analysis

The execution of the complete set 24 of combinations as represented in Figure 1 takes 16 minutes on average. Although this may seem slow, it is encouraging given the complexity of the setup and the inevitable overhead generated by the network data transfers, logging activities of the supporting platform *etc.*

We have run series of evaluations over groups of 30 or 40 images at a time, using the UNLV dataset [9] and

[1] http://www.imagemagick.org/script/convert.php

[2] http://www.gnu.org/software/ocrad/

[3] We deliberately offered acces to an older version of Tesseract, to allow contributions and comparisons with newer versions

[4] http://www.lrde.epita.fr

[5] http://cactus.nci.nih.gov/

[6] Available for download at http://dae.cse.lehigh.edu/DAE/sites/default/files/ICDARContestTotal.t2flow

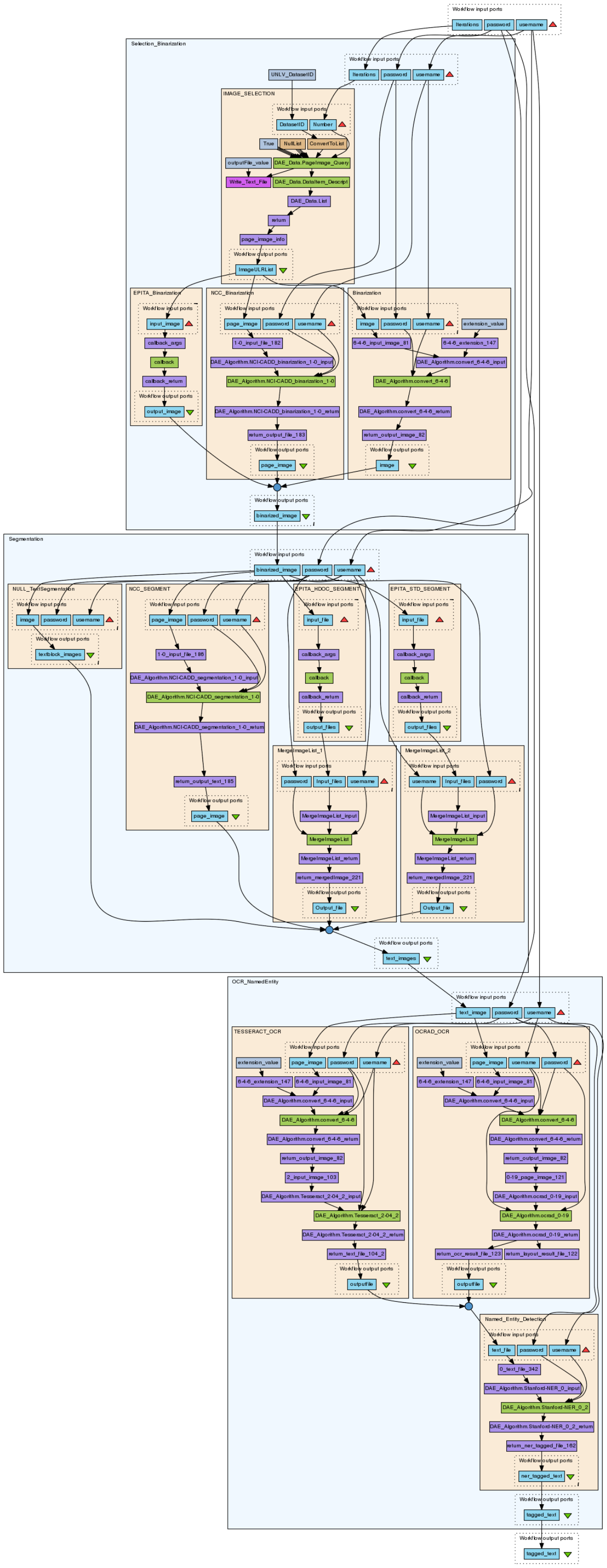

Figure 1. Full contest pipeline with contributed algorithms.

| | |
|---|---|
| 1 | EPITA_BI – NULL_SEG – TESSERACT |
| 2 | EPITA_BI – NULL_SEG – Ocrad |
| 3 | EPITA_BI – NCI_SEG – TESSERACT |
| 4 | EPITA_BI – NCI_SEG – Ocrad |
| 5 | EPITA_BI – EPITA_HDOC – TESSERACT |
| 6 | EPITA_BI – EPITA_HDOC – Ocrad |
| 7 | EPITA_BI – EPITA_STD – TESSERACT |
| 8 | EPITA_BI – EPITA_STD – Ocrad |
| 9 | BI – NULL_SEG – TESSERACT |
| 10 | BI – NULL_SEG – Ocrad |
| 11 | BI – NCI_SEG – TESSERACT |
| 12 | BI – NCI_SEG – Ocrad |
| 13 | BI – EPITA_HDOC – TESSERACT |
| 14 | BI – EPITA_HDOC – Ocrad |
| 15 | BI – EPITA_STD – TESSERACT |
| 16 | BI – EPITA_STD – Ocrad |
| 17 | NCI_BI – NULL_SEG – TESSERACT |
| 18 | NCI_BI – NULL_SEG – Ocrad |
| 19 | NCI_BI – NCI_SEG – TESSERACT |
| 20 | NCI_BI – NCI_SEG – Ocrad |
| 21 | NCI_BI – EPITA_HDOC – TESSERACT |
| 22 | NCI_BI – EPITA_HDOC – Ocrad |
| 23 | NCI_BI – EPITA_STD – TESSERACT |
| 24 | NCI_BI – EPITA_STD – Ocrad |

Table I
CORRESPONDING INDEX LABELS AND EXECUTION PATHS

its associated text transcription as ground-truth. We also conducted evaluations using a new paradigm capable of ovecoming the absence of formalized ground-truth [10], but results were not yet sufficiently compiled at the time of this writing.

Figure 2 shows the overall F-Measure for all 24 combinations and the ranking obtained by ordering these results by decreasing order of F value. The numbering of the 24 combinations corresponds to a depth-first traversal of the execution tree depicted in Figure 1 with one notable inversion at the first level. Table I gives the full correspondance of indexes and execution paths.

The top five best ranked combinations are:

| | |
|---|---|
| 3 | EPITA_BI – NCI_SEG – TESSERACT |
| 17 | NCI_BI – NULL_SEG – TESSERACT |
| 19 | NCI_BI – NCI_SEG – TESSERACT |
| 21 | NCI_BI – EPITA_HDOC – TESSERACT |
| 23 | NCI_BI – EPITA_STD – TESSERACT |

From this finding and from the serrated shape of the graph we can draw the following conclusions.

- Tesseract significantly outperforms Ocrad and provides consistently higher overall performance. This is clearly illustrated in Figure 3, where we have projected the overall results of Figure 2 onto the two OCR dimensions.
- The four overall best ranked combinations involve the NCI_BI binarization program.

These results are confirmed when projecting the overall results of Figure 2 onto either the binarization dimensions (Figure 4) or the segmentation dimensions (Figure 5).

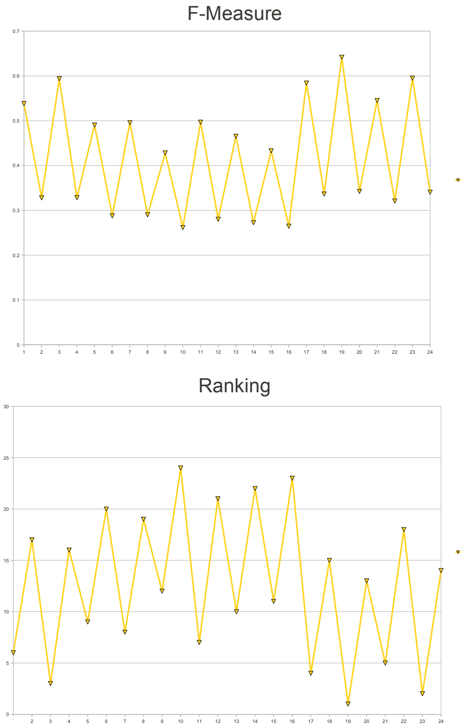

Figure 2. Overall F-Measure and Ranking results for all possible combinations

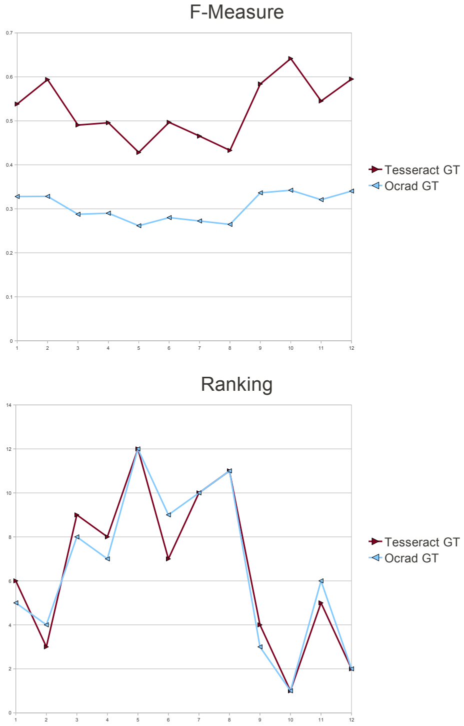

Figure 3. OCR compared F-Measure and Ranking results for all possible combinations

The NCI contributed algorithms consistently outperform the EPITA contributed algorithms, and the impact on the F-Measure is higher for the NCI binarization algorithm.

## VI. Conclusion and Further Work

The first conclusion of this contest is that the paradigm developed in [1], [2] is effective and usable: user contributions were integrated seamlessly, whether they were hosted within or without the DAE framework, regardless of their operating environment. Since it explicitely stores all intermediate results and conditions under which they were obtained, it opens the door to a wider range of uses, contributing to making tracability, reproducibility and comparison of experimental results easier.

Second, and more related to the evaluation of the contributions, the NCI Binarization algorithm is the contributed algorithm that had the most significant positive impact to the end-to-end application. It can be argued that the Tesseract OCR had an even bigger impact, but it was not a competing algorithm. It is also noteworthy to observe that for the Ocrad OCR binarization nor segmentation seem to have any impact on the final result. This might be related to the fact that the software possibly includes some filtering and layout functions, and would be interesting to investigate further.

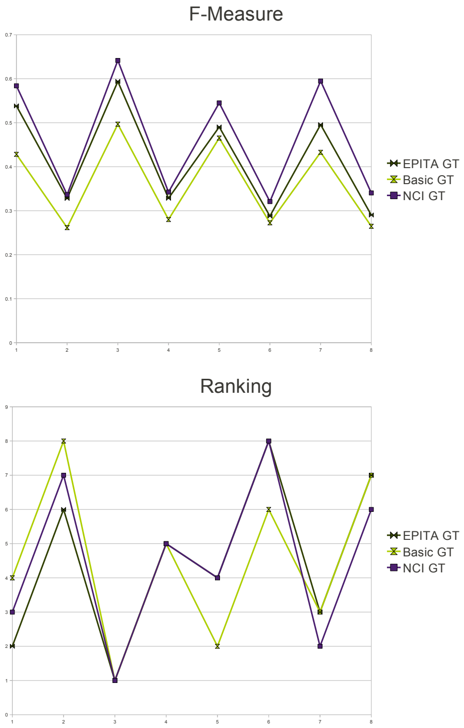

Figure 4. Binarization compared F-Measure and Ranking results for all possible combinations

It is also important to stress that the full contest can be replayed and verified either by downloading the associated

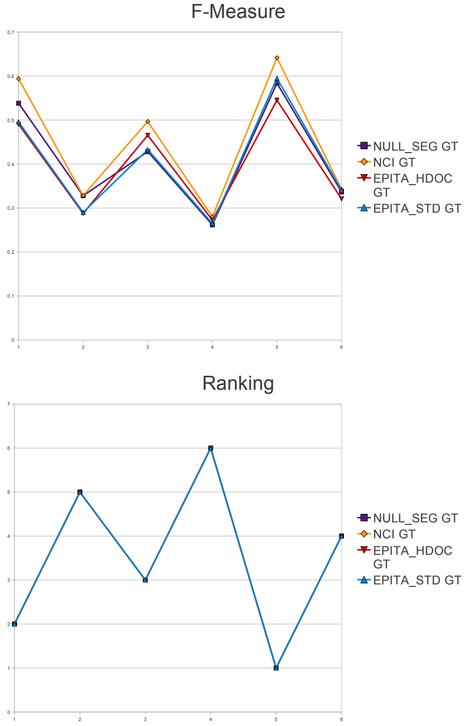

Figure 5. Segmentation compared F-Measure and Ranking results for all possible combinations

Taverna workflow files, or by directly querying the hosting server http://dae.cse.lehigh.edu where all execution information is available. As an example, the following SQL query will give all input and ouptut data that has been produced, and by which algorithms. Due to space constraints, this query has been slightly simplified.

```
select dit_in.id data_item_in, dit_out.id
   data_item_out, alg.id algo_id from
algorithm_run arun
join algorithm_run_of arof on
  arof.algorithm_run_id = arun.id
join algorithm_run_input arin on
          arin.algorithm_run_id = arun.id
join data_item dit_in on
  dit_in.id = arin.data_item_id
join algorithm alg on
  alg.id = arof.algorithm_id
join algorithm_run_output arout on
          arout.algorithm_run_id = arun.id
join data_item dit_out on
  dit_out.id = arout.data_item_id
where alg.id in (182,183,143,162,125,222) and
arun.start_time >= to_date('01/jun/2011',
'dd/mm/yyyy') and
arun.start_time < to_date('01/jul/2011',
'dd/mm/yyyy')
```

Full queries can be obtained from http://dae.cse.lehigh.edu/DAE/?q=node/60.

*Acknowledgements*

Bart Lamiroy was a visiting scientist at Lehigh University in 2010-2011 at the Computer Science and Engineering Department. This work was funded through DARPA IPTO via Raytheon BBN Technologies.

## References

[1] B. Lamiroy and D. Lopresti, P., "An Open Architecture for End-to-End Document Analysis Benchmarking," in *11th International Conference on Document Analysis and Recognition - ICDAR 2011*. Beijing, China: International Association for Pattern Recognition, Sep. 2011.

[2] B. Lamiroy, D. Lopresti, H. Korth, and J. Heflin, "How Carefully Designed Open Resource Sharing Can Help and Expand Document Analysis Research," in *Document Recognition and Retrieval XVIII - DRR 2011*, G. Agam and C. Viard-Gaudin, Eds., vol. 7874, SPIE. San Francisco, United States: SPIE, Jan. 2011, iSBN : 9780819484116.

[3] T. Oinn, M. Greenwood, M. Addis, N. Alpdemir, J. Ferris, K. Glover, C. Goble, A. Goderis, D. Hull, D. Marvin, P. Li, P. Lord, M. Pocock, M. Senger, R. Stevens, A. Wipat, and C. Wroe, "Taverna: lessons in creating a workflow environment for the life sciences," *Concurrency and Computation: Practice and Experience*, vol. 18, no. 10, pp. 1067–1100, August 2006.

[4] R. Smith, "An overview of the tesseract ocr engine," in *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 2007, pp. 629–633.

[5] J. R. Finkel, T. Grenager, and C. D. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *ACL*. The Association for Computer Linguistics, 2005.

[6] G. Lazzara, R. Levillain, Th. Géraud, Y. Jacquelet, J. Marquegnies, and A. Crépin-Leblond, "The scribo module of the olena platform: a free software framework for document image analysis," in *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR)*. Beijing, China: International Association for Pattern Recognition (IAPR), Sep. 2011.

[7] J. J. Sauvola and M. Pietikäinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, 2000.

[8] I. V. Filippov and M. C. Nicklaus, "Optical structure recognition software to recover chemical information: OSRA, an open source solution," *Journal of Chemical Information and Modeling*, vol. 49, no. 3, pp. 740–743, 2009.

[9] T. A. Nartker, R. B. Bradford, and B. A. Cerny, "A preliminary report on UNLV/GT1: A database for ground-truth testing in document analysis and character recognition," in *Proceedings of the First Symposium on Document Analysis and Information Retrieval SDAIR 92*, University of Nevada, Las Vegas, March 1992, pp. 300–315.

[10] B. Lamiroy and T. Sun, "Precision and recall without ground truth," in *Ninth IAPR International Workshop on Graphics RECognition - GREC 2011*, Seoul, Korea, 2011.