Objectives

- Digital transformation of engineering documents
- Extraction of textual entities on large and unstructured documents
- Short text contextualisation and correct OCR predictions of Tags.

Related works

**Post-OCR Correction**


**Text Detection**


Contribution

**Detection**:  
- Based on the FCN model EAST with pre-trained Resnet V1.  
- NMS part is removed by our own filtering method.  
- Adaptation of the system to the use case

**Post-OCR Correction**:  
- Clustering of Tags using Affinity Propagation  
- Identification of potential errors made by the OCR by analyzing the tag structure  
- Proposal to correct misrecognized characters

Experiments

- Industrial data are sensitive and cannot be shared, so we have tested on our own dataset  
- Calculation of the Precision, Recall & F1 Score for the detection module  
- The Affinity Propagation clusters are the tags of the plan “shot by shot”  
- Difference between the WER of the OCR output and the post-OCR Correction: 7% (75% OCR–82% post-

Conclusion & future works

- Focus on graphic symbols to help the post-OCR correction  
- Association of symbolic and textual entities  
- Improve initial step of text recognition by OCR competition

Evaluation of the detection system with our own dataset