Intelligent Document Processing with Small and Relevant Training Dataset

Lina NICOLAIEFF, ***Mohamed KANDI***,
Younes ZEGAOUI, *Christophe BORTOLASO*

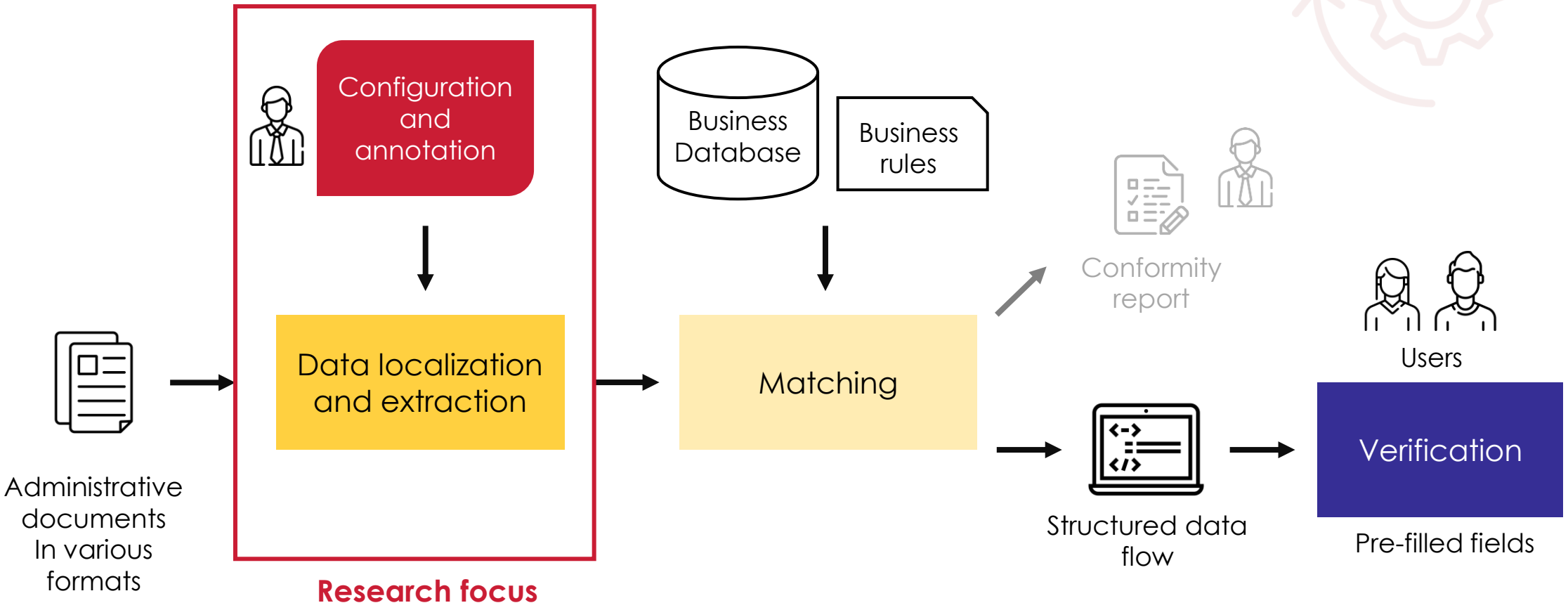
FirstName.LastName@berger-levrault.com
Berger-levrault, FRANCE

APIA 2022, Saint-Étienne, France

Plan of the presentation

- Context and problem
- State of the art
- methodology & System description
- Experiments
- Conclusion and perspectives
- Q&As

Context

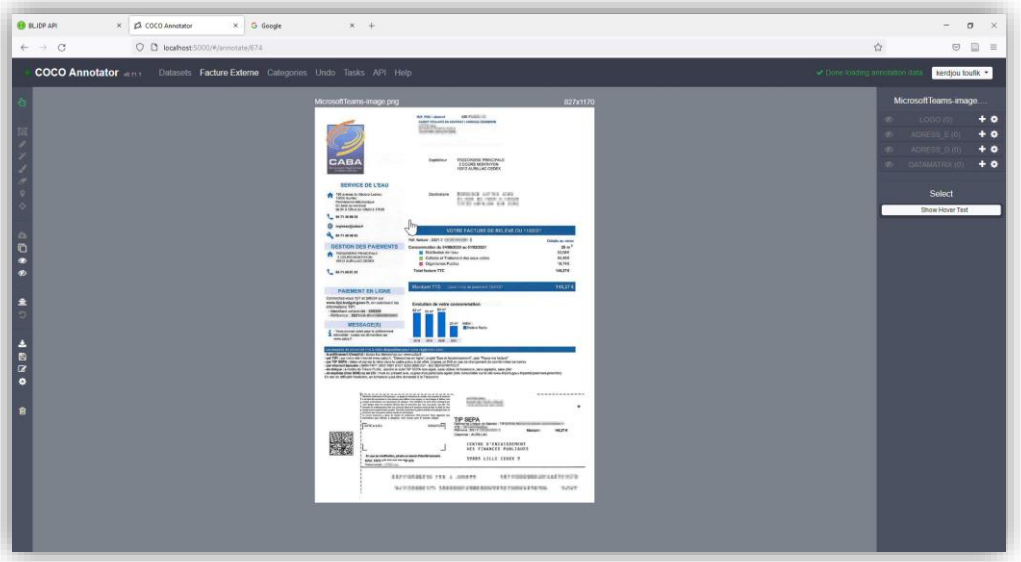


Generic process of a typical IDP approach

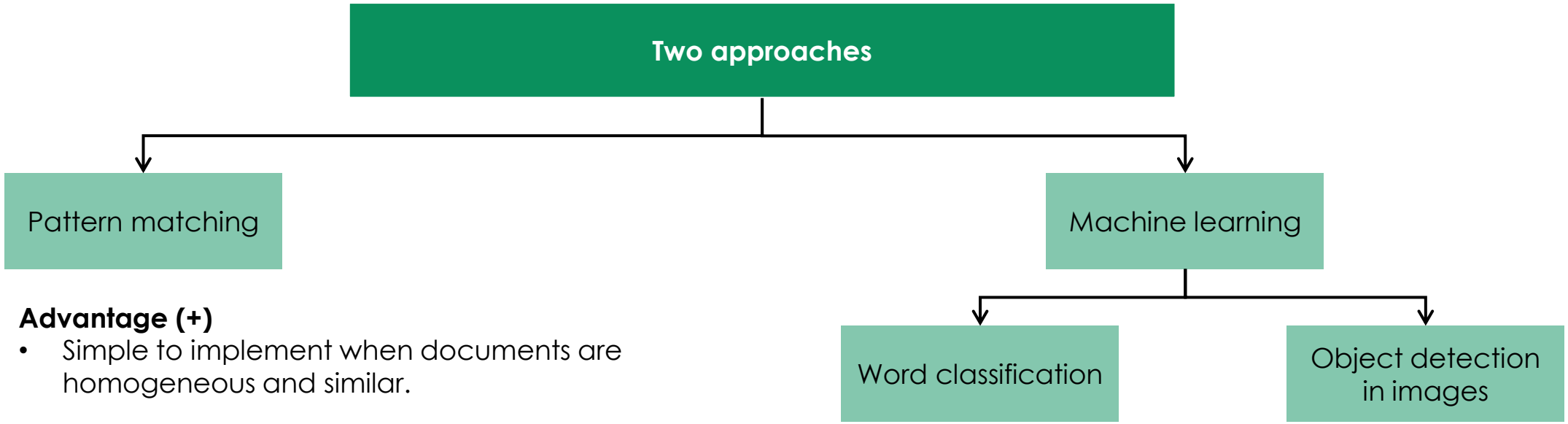
Problem

- Annotation is a **tedious** and **repetitive** task done regularly when new document formats are introduced.

- How to select a **small** and **relevant** subset of unstructured document to annotate in order to **reduce** data annotation **effort** ?



State of the art



Advantage (+)
- Simple to implement when documents are homogeneous and similar.

Disadvantages (-)
- Does not generalize,
- Maintenance requires time and expertise.

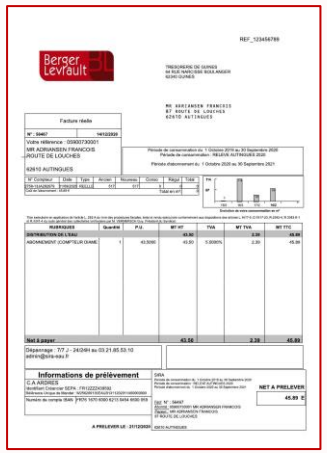
Advantage (+)
- Can be generalized for many document models.

Disadvantages (-)
- The need for a set of annotated examples.

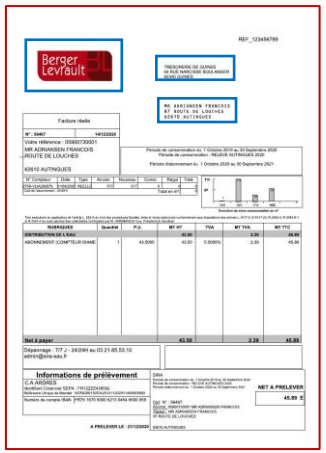
Information Extraction from Unstructured Documents

[1] R. B. Palm, F. Laws, and O. Winther, “Attend, copy, parse end-to-end information extraction from documents,” in 2019 International Conference on Document Analysis and Recognition (ICDAR). IEEE, 2019, pp. 329–336.

Methodology & system description



image



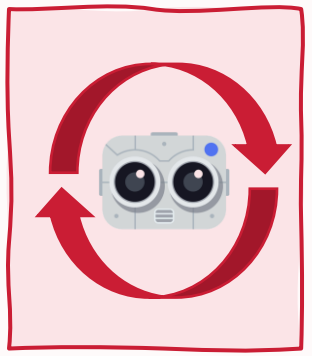
annotation

```
411 149 587 193 ADRESS_E
408 254 573 312 ADRESS_D
17 997 114 1095 LOGO
```

<(x, y, h, w, **class**)>

```
{"image_id":0,
"file_name":"images\/facture_0.jpg",
"height":1170,
"width":827,
"annotations":[{"bbox":[103,87,145,67],
                "bbox_mode":1,"category_id":"0"},
                {"bbox":[474,252,123,48],
                "bbox_mode":1,
                "category_id":"1"}]}
                ...
```

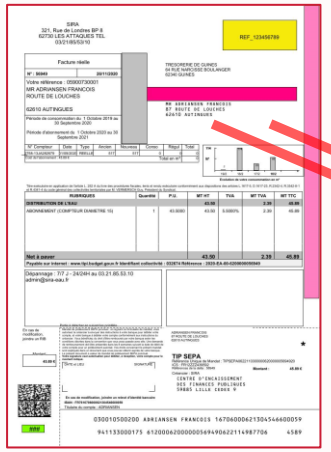
Coco format



Train

Step 1 : Model training

Methodology & system description



PyTesseract

Optical Character Recognition (OCR)

```
MR xxxxx xxxxx
xx ROUTE DE LOUCHES
62610 AUTINGUES
```

New document

Prediction : Object localization in image

<(x, y, h, w, **class**)>

Text extraction from image

Control functions : content & position

<(x, y, h, w, **class**, **text**)>

Step 2 : inference

Methodology & system description

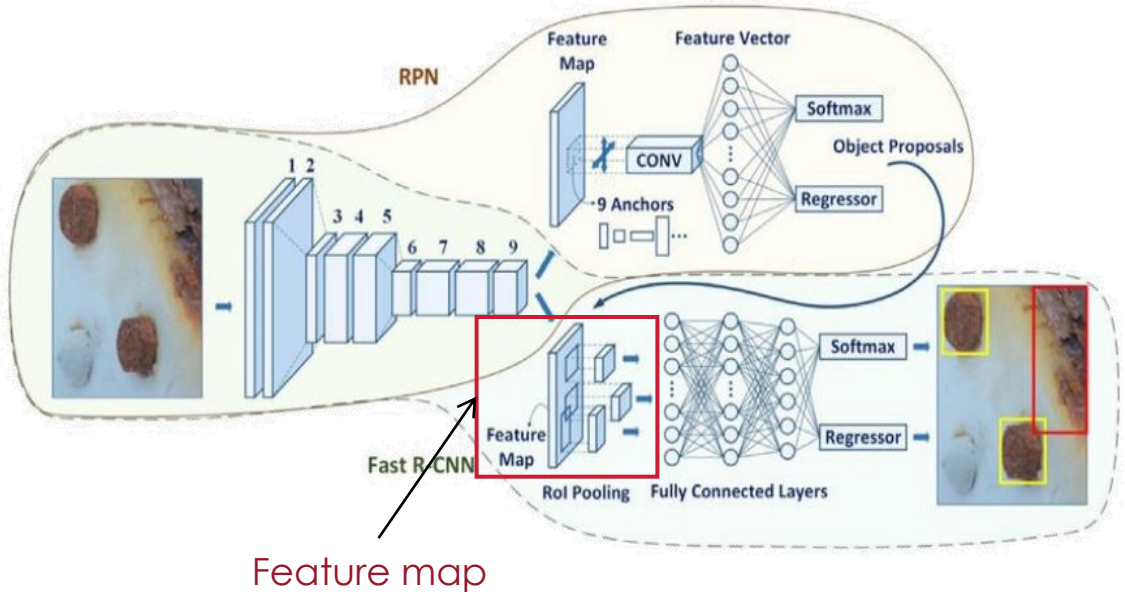
Objet detection model

Learning model used

Faster RCNN architecture (CNN + feature map)

Pre-trained on the COCO dataset :

- 121,408 pictures
- 888,331 annotated objects (box)
- 80 labels



Feature map

Architecture of the model

Evaluation metric : **Average Precision (AP)**

$$IoU = \frac{\text{Area overlap}}{\text{Area union}}$$

$$AP_{\propto} = \int_0^1 p(r)\ dr$$

α : IoU threshold

$$AP = \frac{1}{n} \sum_{\propto = 0,5}^{0,95} AP_{\propto}$$

Average AP calculated on several thresholds

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, “Microsoft coco: Common objects in context,” in European Conference on Computer Vision, 2014, pp. 740–755.

Experiments

Experiment protocol

Evaluate the impact of the number of documents in the training set on the prediction accuracy.

Object to be predicted by the model :

- ✓ Recipient address (0)
- ✓ **Sender address (1)**
- ✓ Logo (2)
- ✓ Datamatrix (3)

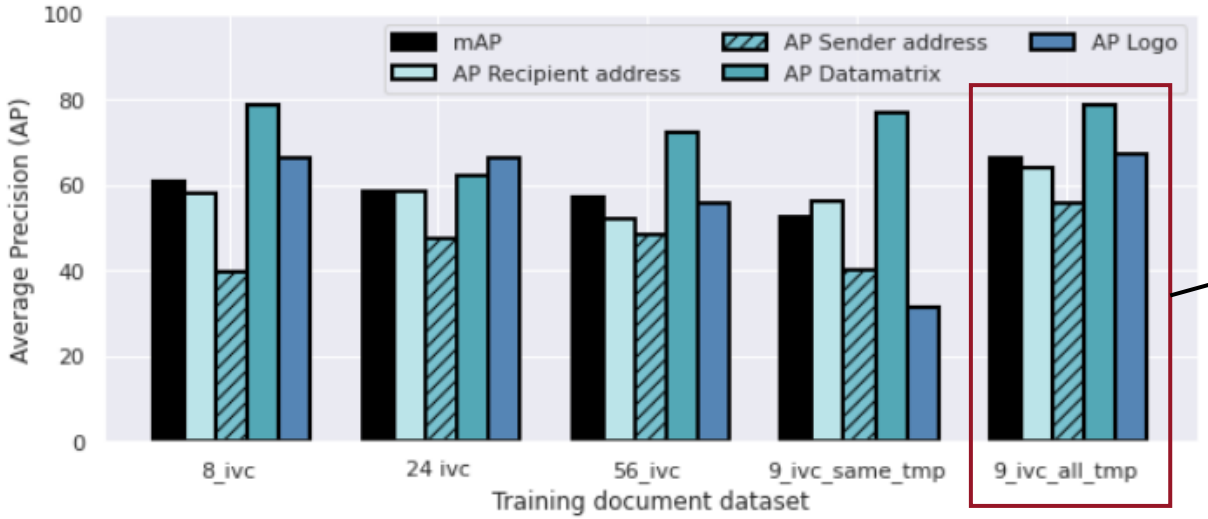
Training datasets :

Evaluate the impact of the number of documents :

- 1 template - 8 documents
- 8 templates - 8, 24, 56 documents
- 9 templates - 9 documents

Experiments

Results



Best prediction score

Training set	mAP	AP Recipient address	AP Sender address	AP Datamatrix	AP Logo
8_ivc	**60.789**	57.954	39.953	78.620	66.627
24_ivc	**58.634**	58.428	47.734	62.111	66.264
56_ivc	**57.267**	52.315	48.406	72.444	55.901
9_ivc_same_tmp	**52.766**	56.230	40.383	76.818	31.634
9_ivc_all_tmp	**66.486**	64.025	55.696	78.742	67.479

TABLE I: Average Precision (AP) detailed results

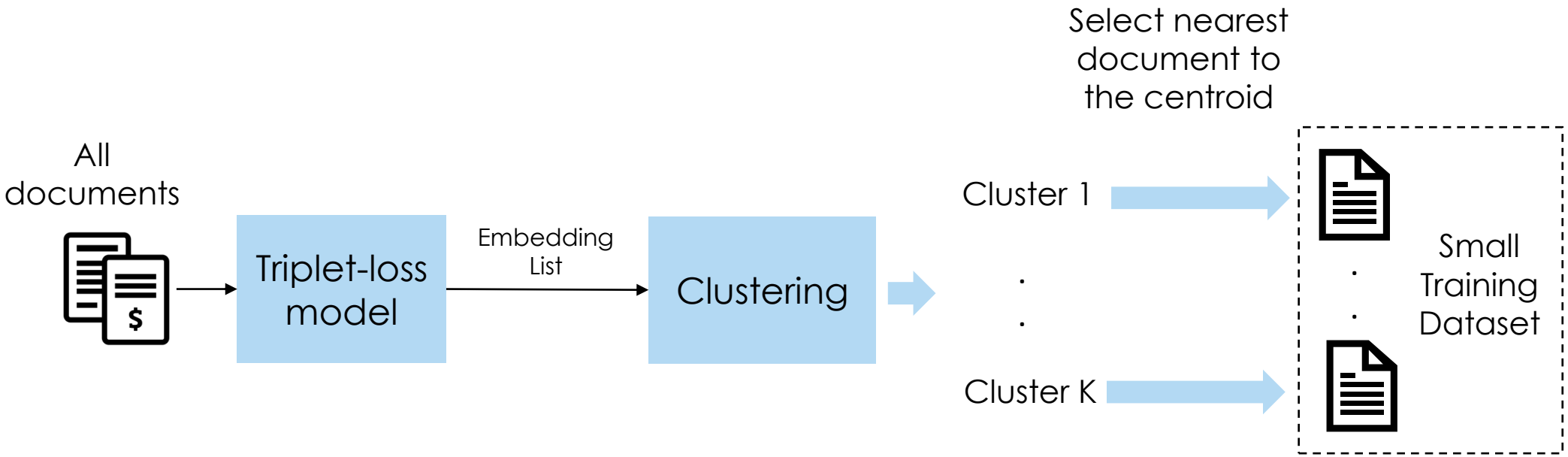
Training set	Recipient add	Sender add	Datamatrix	Logo
8_ivc	8	8	6	4
24_ivc	24	24	19	12
56_ivc	56	56	36	38
9_ivc_same_tmp	9	9	5	6
9_ivc_all_tmp	9	9	4	9

TABLE II: Number of invoices containing each object in the Training sets

AP model results for each dataset

Methodology & system description

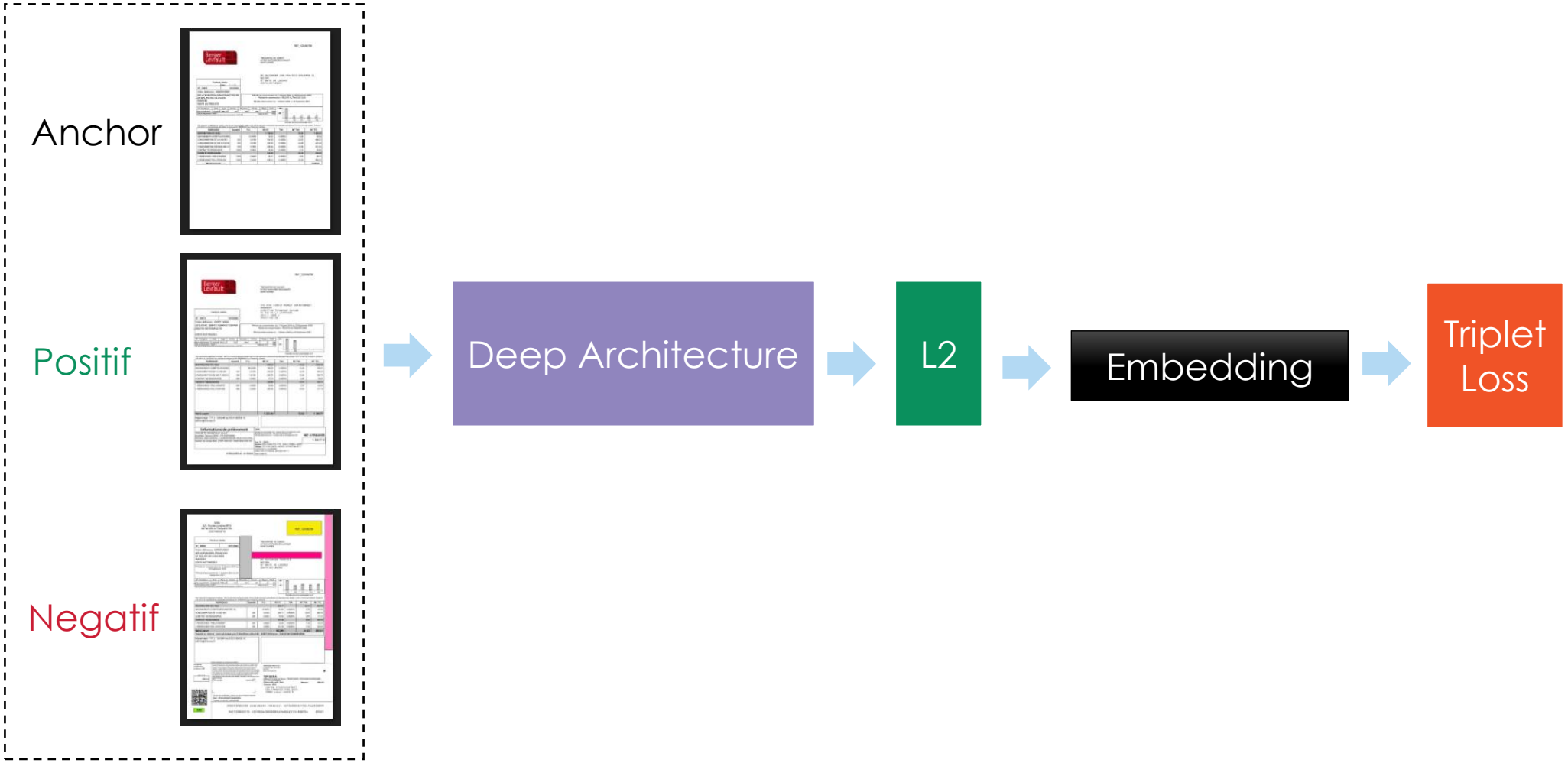
Best training candidate selection



E. Hoffer and N. Ailon, “Deep metric learning using triplet network,”in Similarity-Based Pattern Recognition, A. Feragen, M. Pelillo, and M. Loog, Eds. Cham: Springer International Publishing, 2015, pp. 84–92.

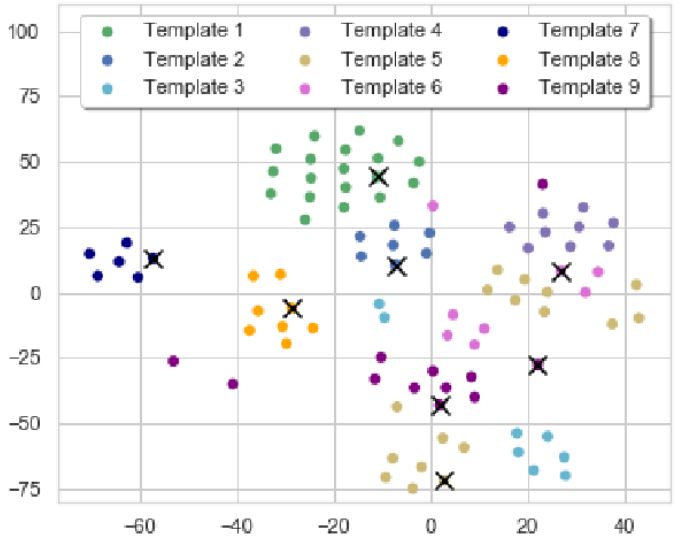
Methodology & system description

Best training candidate selection

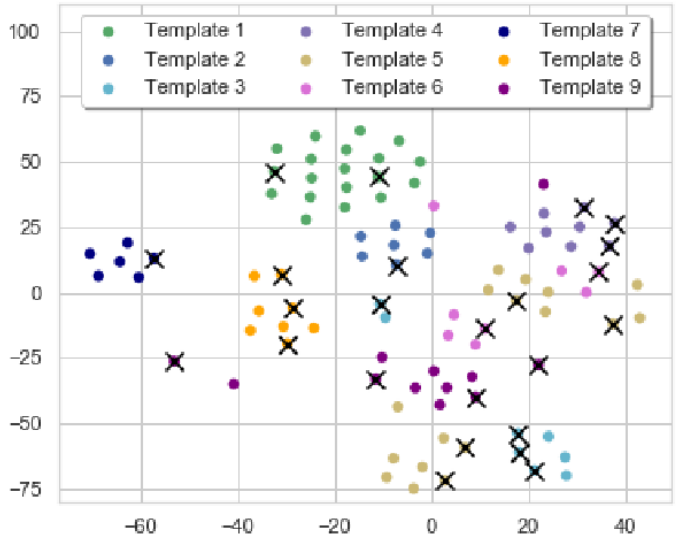


Experiments

A projection of document embeddings

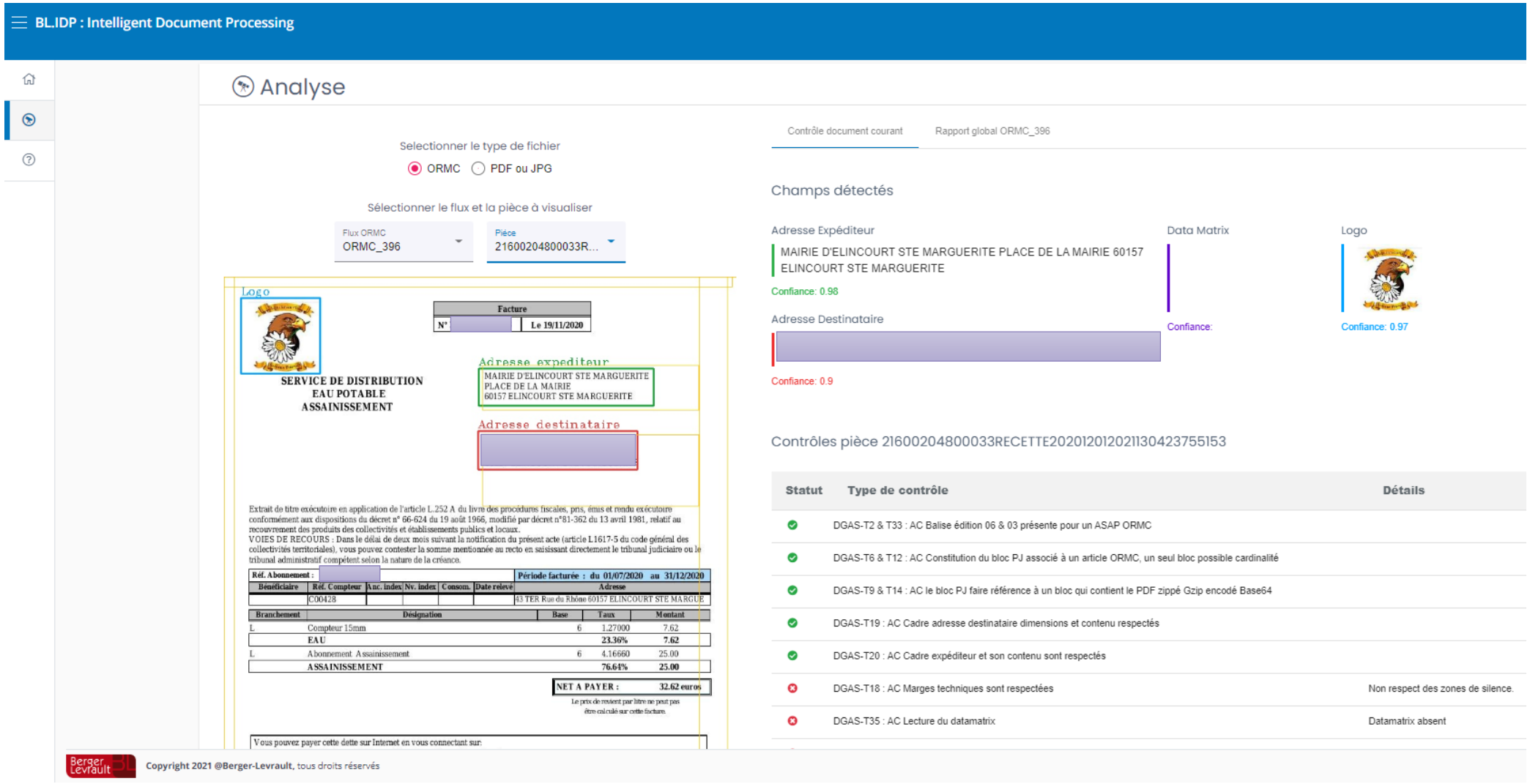


(a) Nearest document to the centroid - Euclidean distance - 8 selected documents



(b) Nearest document to the centroid - Euclidean distance - 24 selected documents

Prototype


BL.IDP : Intelligent Document Processing
Analyse
Selectionner le type de fichier
ORMC
PDF ou JPG
Sélectionner le flux et la pièce à visualiser
Flux ORMC
ORMC_396
Pièce
21600204800033R...
Logo
Facture
N°
Le 19/11/2020
SERVICE DE DISTRIBUTION
EAU POTABLE
ASSAINISSEMENT
Adresse expediteur
MAIRIE D'ELINCOURT STE MARGUERITE
PLACE DE LA MAIRIE
60157 ELINCOURT STE MARGUERITE
Adresse destinataire
Extrait de titre exécutoire en application de l'article L.252 A du livre des procédures fiscales, pris, émis et rendu exécutoire conformément aux dispositions du décret n° 66-624 du 19 août 1966, modifié par décret n°81-362 du 13 avril 1981, relatif au recouvrement des produits des collectivités et établissements publics et locaux.
VOIES DE RECOURS : Dans le délai de deux mois suivant la notification du présent acte (article L1617-5 du code général des collectivités territoriales), vous pouvez contester la somme mentionnée au recto en saisissant directement le tribunal judiciaire ou le tribunal administratif compétent selon la nature de la créance.
Réf. Abonnement :
Période facturée : du 01/07/2020 au 31/12/2020
Bénéficiaire
Réf. Compteur
Anc. index
Nv. index
Consom.
Date relevé
Adresse
C00428
43 TER Rue du Rhône 60157 ELINCOURT STE MARGUE
Branchement
Désignation
Base
Taux
Montant
L
Compteur 15mm
6
1.27000
7.62
EAU
23.36%
7.62
L
Abonnement Assainissement
6
4.16660
25.00
ASSAINISSEMENT
76.64%
25.00
NET A PAYER :
32.62 euros
Le prix de revient par litre ne peut pas être calculé sur cette facture.
Vous pouvez payer cette dette sur Internet en vous connectant sur:
Contrôle document courant
Rapport global ORMC_396
Champs détectés
Adresse Expéditeur
MAIRIE D'ELINCOURT STE MARGUERITE PLACE DE LA MAIRIE 60157 ELINCOURT STE MARGUERITE
Confiance: 0.98
Data Matrix
Confiance:
Logo
Confiance: 0.97
Adresse Destinataire
Confiance: 0.9
Contrôles pièce 21600204800033RECETTE20201201202113042375515 3
Statut
Type de contrôle
Détails
DGAS-T2 & T33 : AC Balise édition 06 & 03 présente pour un ASAP ORMC
DGAS-T6 & T12 : AC Constitution du bloc PJ associé à un article ORMC, un seul bloc possible cardinalité
DGAS-T9 & T14 : AC le bloc PJ faire référence à un bloc qui contient le PDF zippé Gzip encodé Base64
DGAS-T19 : AC Cadre adresse destinataire dimensions et contenu respectés
DGAS-T20 : AC Cadre expéditeur et son contenu sont respectés
DGAS-T18 : AC Marges techniques sont respectées
Non respect des zones de silence.
DGAS-T35 : AC Lecture du datamatrix
Datamatrix absent
Copyright 2021 @Berger-Levrault, tous droits réservés

Conclusion and perspectives

- In this work, we have shown that the Triplet-loss based model combined with clustering can be used to select a subset of relevant documents to annotate and train a Faster R-CNN model.
- In future work :
  - Conduct experiments on a larger number of templates
  - Expand our work by designing new experiments :

    1. *unify the Triplet-loss model with the CNN detector model by making them share some of their features,*
    2. *compare the regular Triplet-loss + k-means model with a unified deep embedding clustering (DEC) approach,*
    3. *going further in the Few-shot learning direction by leveraging existing methods such as the matching networks to help our model get the most information from our dataset during training.*

Thank you for your attention
Discussion !

berger-levrault.com