HANDWRITTEN/PRINTED TEXT SEPARATION USING PSEUDO-LINES FOR CONTEXTUAL RE-LABELING

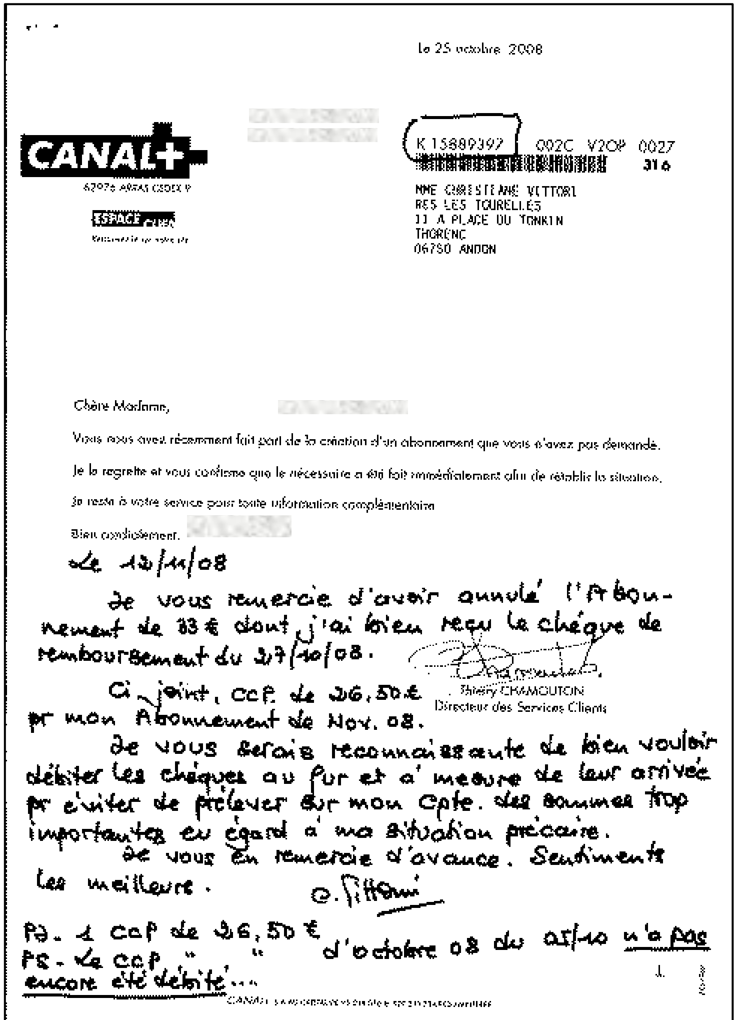
By:

Ahmad Montaser Awal
Abdel Belaïd
Vincent Poulain d'Andecy

Context

- Administrative documents are
  - Noisy
  - Annotated…
- Separation of scripts in administrative documents
  - Annotation extraction
  - Sending each script to a specialized system
  - Noise removal

CONTEXT


le 25 octobre 2008

CANAL+
62976 ARRAS CEDEX 9

K 15889392 092C V2OP 0027 316

MME CHRISTIANE VITTORI
RES LES TOURELLES
11 A PLACE DU TONKIN
THORENC
06750 ANDON

Chère Madame,

Vous nous avez récemment fait part de la création d'un abonnement que vous n'avez pas demandé.

Je le regrette et vous confirme que le nécessaire a été fait immédiatement afin de rétablir la situation.

Je reste à votre service pour toute information complémentaire.

Bien cordialement,

Thierry CHAMOUTON
Directeur des Services Clients

le 13/11/08

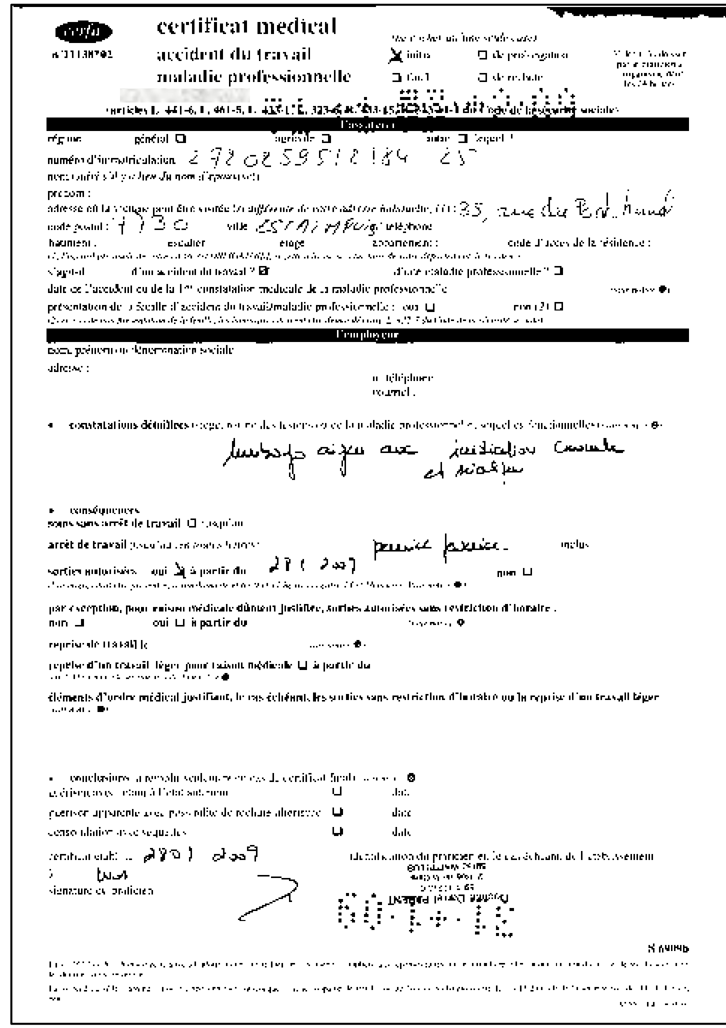
Je vous remercie d'avoir annulé l'Abonnement de 33 € dont j'ai bien reçu le chèque de remboursement du 27/10/08.

Ci-joint, ccp de 26,50 € pr mon Abonnement de Nov. 08.

Je vous serais reconnaissante de bien vouloir débiter les chèques au fur et à mesure de leur arrivée pr éviter de prélever sur mon cpte. des sommes trop importantes eu égard à ma situation précaire. Je vous en remercie d'avance. Sentiments les meilleurs.

C. Vittori

PS. 1 ccp de 26,50 € d'octobre 08 du 25/10 n'a pas encore été débité...


certificat médical
accident du travail
maladie professionnelle

State of the art

- Printed/handwritten text separation systems share the main steps
  - Preprocessing

    Removing very small/large connected components
  - Document segmentation

    Segment the document into basic units
  - Classification

    Assign each unit to a text class
  - Contextual re-labeling

    Correct classification errors using neighborhood information

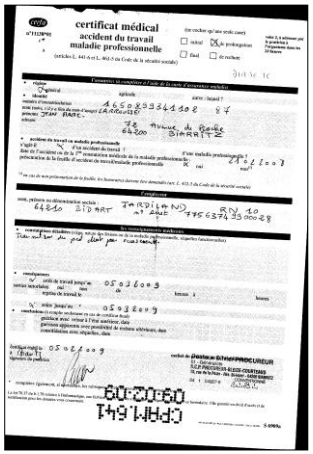
State of the art
Document segmentation

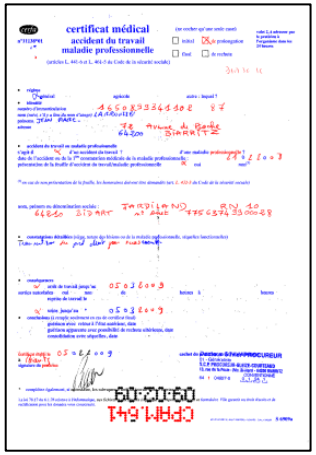
- Text line level (Pal et al. 2001)(Kavallieratou et al. 2004)
  - Lines are assumed to be homogeneous (mono-class)
  - Segmentation using the horizontal projection profiles
- Word level
  - Grouping connected components to approximate words
  - Distance based (Zheng et al. 2004) (Shetty et al. 2007)
  - Morphological operations (Peng et al. 2011) (Zagoris et al. 2014)
- Character level (Fan et al. 1998)
  - Non-cursive scripts (Chinese documents)
  - X-Y cut algorithm

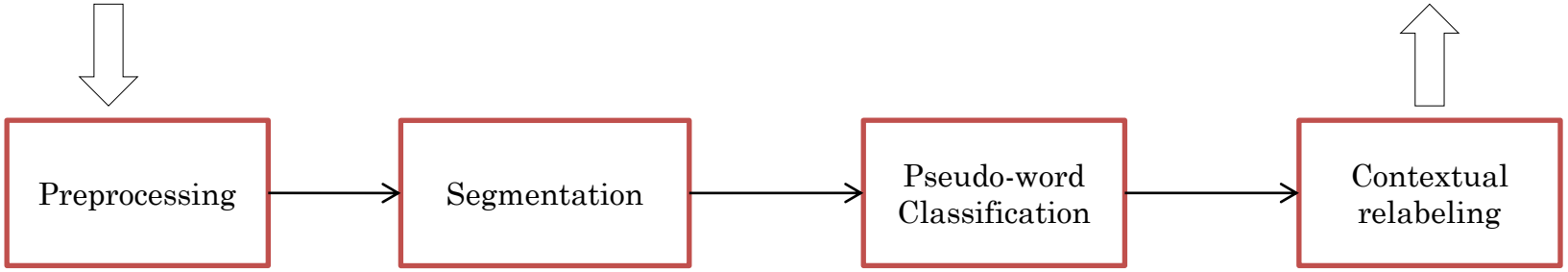
State of the art
Contextual Re-labeling

- Step1: Define the neighborhood of a given word
  - 4 Nearest Neigbors (Peng et al. 2013) (Zheng et al. 2007)
  - 6 Nearest Neighbors (Shetty et al, 2007)
- Step2: Define criteria to re-label a word based on the labels of its neighborhood
  - Majority voting (kandan et al. 2007)
  - Probabilistic models
    - Markov Random Field (MRF) (Zheng et al. 2007) (Peng et al. 2013)
    - Conditional Random Field (CRF) (Shetty et al. 2007)

PROPOSED SYSTEM OVERVIEW







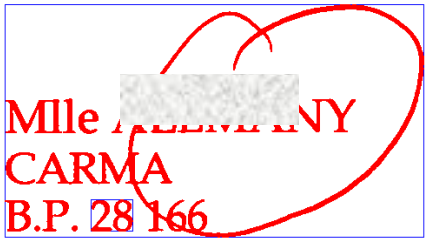
[*] A. Belaïd, K. Santoch and V. Poulain d'Andecy, "Handwritten and Printed Text Separation in Real Document," *Machine Vision Applications*, vol. 2, 2013

SEGMENTATION

- Differently from most of existing works, the document is first segmented into pseudo-lines before being segmented into pseudo-words
- Pseudo-line
  - A set of connected components where:
    - Horizontal distances < d_H
    - Vertical distances < d_V
- Pseudo-word
  - A set of connected components belonging to the same pseudo-line
  - Horizontal distance < *ws* (word spacing distance estimated automatically for each pseudo-line)

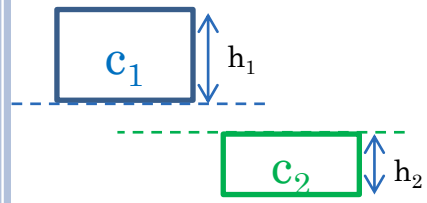
IMPROVED SEGMENTATION – HEURISTIC

- Avoid vertical connection caused by handwritten annotations

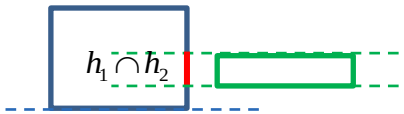


- Use CCs horizontal overlapping

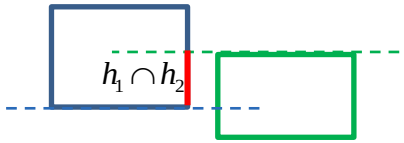
$$o(c_1, c_2) = \frac{h_1 \cap h_2}{\max(h_1, h_2)}$$



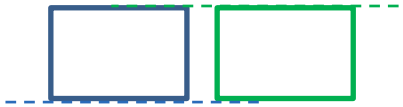
O= 0%



O =30 %

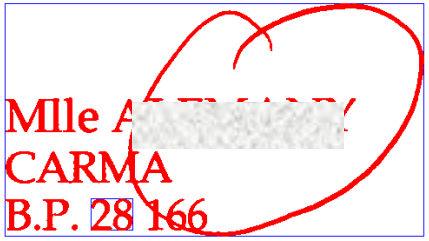


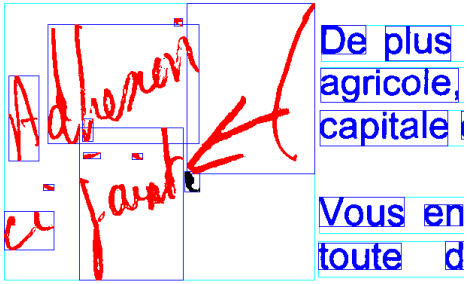
O = 50%

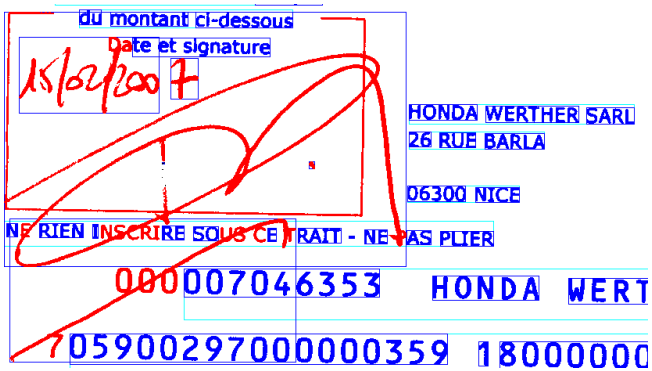


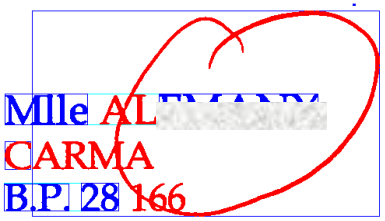
O = 100%

IMPROVED SEGMENTATION – HEURISTIC









Pseudo-words classification

- A pseudo-word is characterized by 137 features
- A multiclass Support vector machines SVM is used to classify a pseudo-word into :
  - Handwritten text
  - Printed text
  - Noise

CONTEXTUAL RELABELING

- Some classification errors could be corrected using contextual neighborhood
- The label of each pseudo-word is updated based on those of its neighbors
- Local neighborhood
  - K nearest neighbors*
  - Confidence propagation *
  - Conditional Random Fields
- Using pseudo-lines
  - Probabilistic model (CRF)
  - Static model

* A. Belaïd, K. Santoch and V. Poulain d'Andecy, "Handwritten and Printed Text Separation in Real Document," *Machine Vision Applications,* vol. 2, 2013

Conditional Random Fields (CRF)

- The separation problem can be modeled by CRF
- According to (Nicolas et al. 2007), the probability of a pseudo-word w is given by:

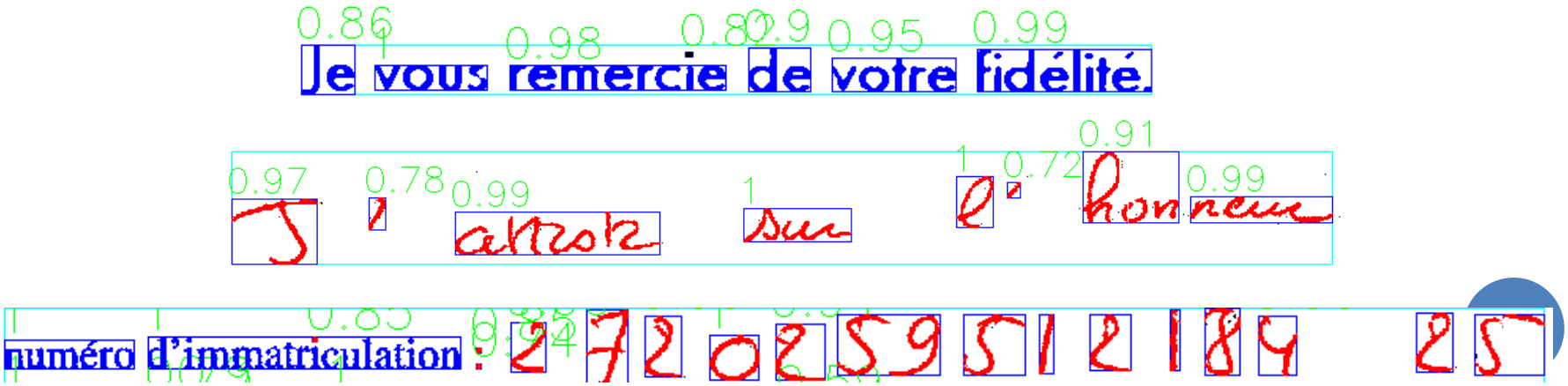
$$P(X_w | Y_L, Y_C) = \lambda_L f_L + \lambda_C f_C$$

Label field — Local features — Contextual features — Local classifier — Contextual classifier

- Contextual features
  - Local classification probabilities of left/right neighbors
  - Structural features extracted from the pseudo-word and each neighbor
    - Height ratio
    - Position ratios
    - Density ratio

Re-labeling using pseudo-lines

- Ideally, a pseudo-line represents a text line of the document
- More than 90% of pseudo-lines contain one type of text (printed or handwritten)
- Pseudo-lines define, implicitly, a global horizontal neighborhood relation between the pseudo-words



RE-LABELING USING PSEUDO-LINES

- The ***dominant class*** C_D in a pseudo-line is the class with the highest cardinality
- In case of equality of cardinalities, the dominant class is the one with highest average confidence of its pseudo-words
- The label of a pseudo-word is updated:
  - Using a CRF model
  - If it verifies the following condition:

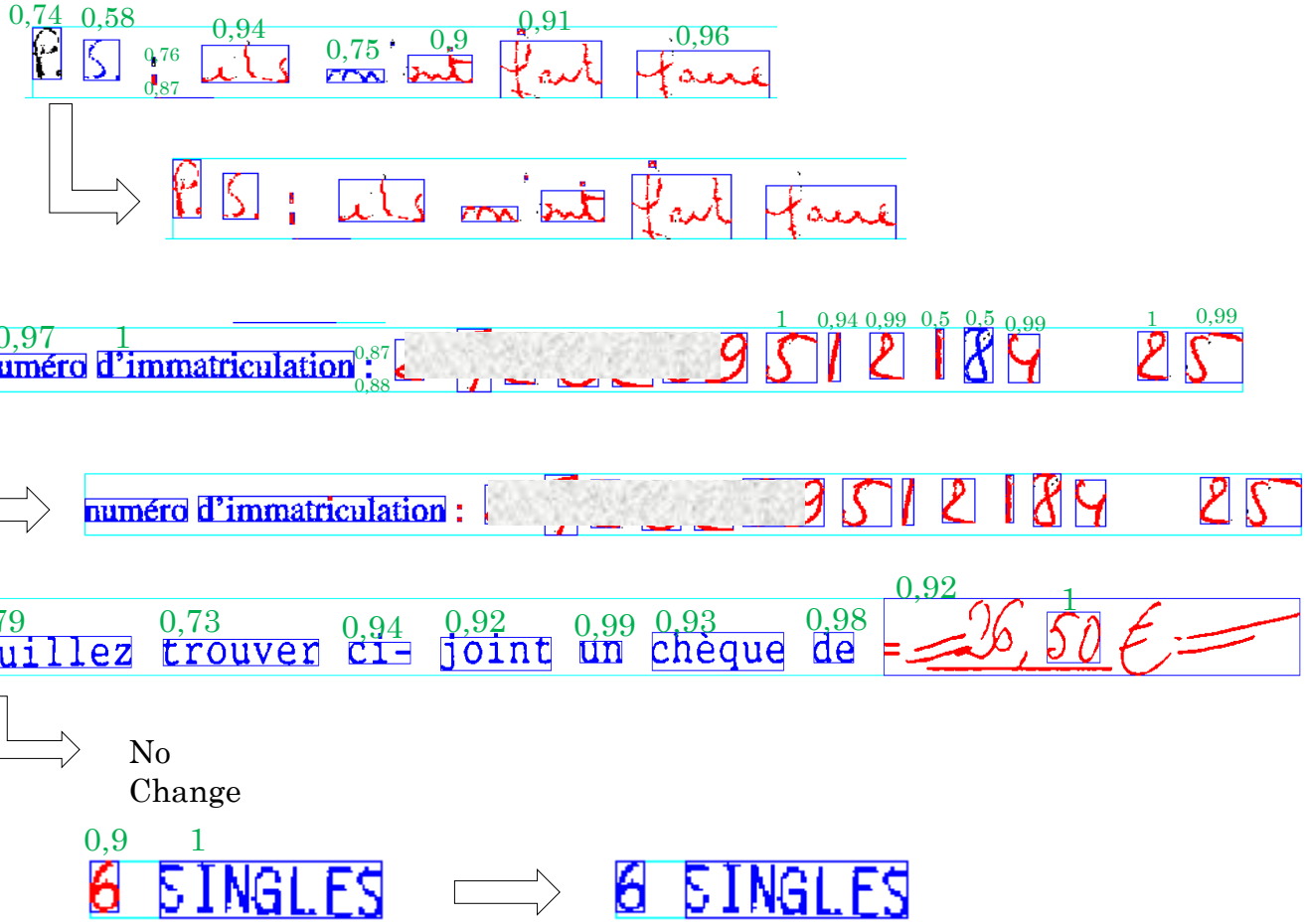
$$(f_i < cf) \vee (|h_i - h_D| < d)$$

Classification Confidence

Certainty factor

Regularity factor

Re-labeling using pseudo-lines Examples



EXPERIMNTATION

- Evaluation
  - Pixel level $$pixRate = \frac{\text{pixels correctly recognised}}{\text{total number of pixels}}$$
  - Pseudo-word level $$pwRate = \frac{\text{pseudo-words correctly recognized}}{\text{total number of pseudo-words}}$$
- Documents
  - Training DB
    - 107 documents (32706 pseudo-words)
      - H: 5888; P: 18078; N: 8740
  - Test DB
    - 202 documents (82142 pseudo-words)
      - H: 11970; P: 43705; N: 25190
  - All documents are labeled at the pixel level

RESULTS (1/2)

	System	H%	P%	N%
Previously proposed system*	Proposed system without contextual re-labeling	97.7	96.5	94.3
	k-NN	95.5	97.5	92.3
	Confidence propagation	97.8	96.6	94.0
New relabeling methods	CRF	98.5	97.1	94.2
	Pseudo-lines (CRF): Probabilistic	98.9	97.5	93.5
	Pseudo-lines: Deterministic	98.3	99.2	87.9
Improved segmentation	Pseudo-lines: Deterministic	99.1	99.2	90.1

* A. Belaïd, K. Santoch and V. Poulain d'Andecy, "Handwritten and Printed Text Separation in Real Document," *Machine Vision Applications,* vol. 2, 2013

RESULTS (2/2)

		pwRate			*pixRate*			
System	Docs	H%	P%	ALL%	H%	P%	N%	ALL%
[kandan et al. 2007]	150	-	-	93.2	-	-	-	-
[Zheng et al. 2004]	94	93.0	98,0	98.1	-	-	-	-
[Peng et al. 2013]	82	93.8	95,7	95.5	-	-	-	-
[Shetty et al. 2007]	27	-	-	-	94.8	98.4	89.8	95.7
[Hamrouni et al. 2014]	32	-	-	-	80.0	92.8	-	90.1
Proposed system	202	97.3	99.5	98.7	99.1	99.2	90.1	96.8

Conclusion and perspectives

- Distance based segmentation is not always enough to obtain 'good' pseudo-words
  - Heuristics could improve and solve some segmentation problems
- A better performance using pseudo-line based contextual relabeling
- A very good performance compared to the state of the art systems
- In future work:
  - Feature selection
  - Ambiguity layer

Thank *you*