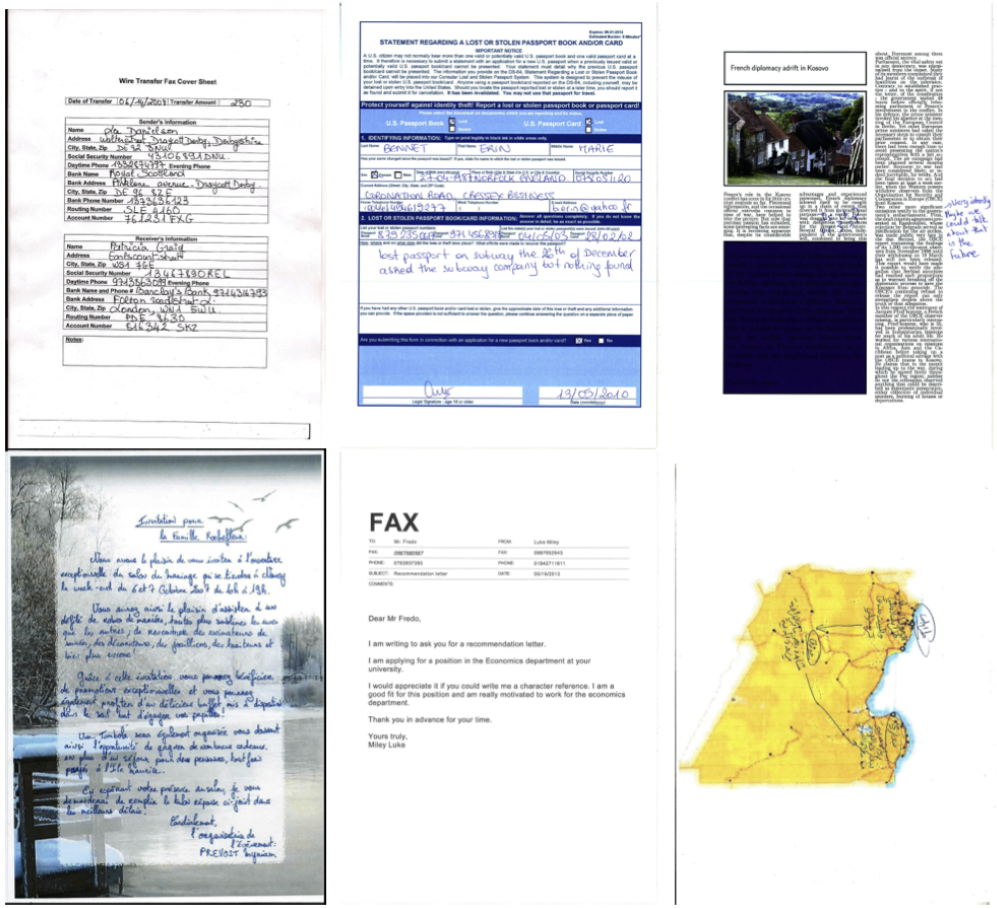
The A2iA Multi-lingual Text Recognition System at the second Maurdor Evaluation

Bastien Moysset, Théodore Bluche, Maxime Knibbe
Mohamed Faouzi Benzeghiba, Ronaldo Messina, Jérôme Louradour
Christopher Kermorvant

www.a2ialab.com

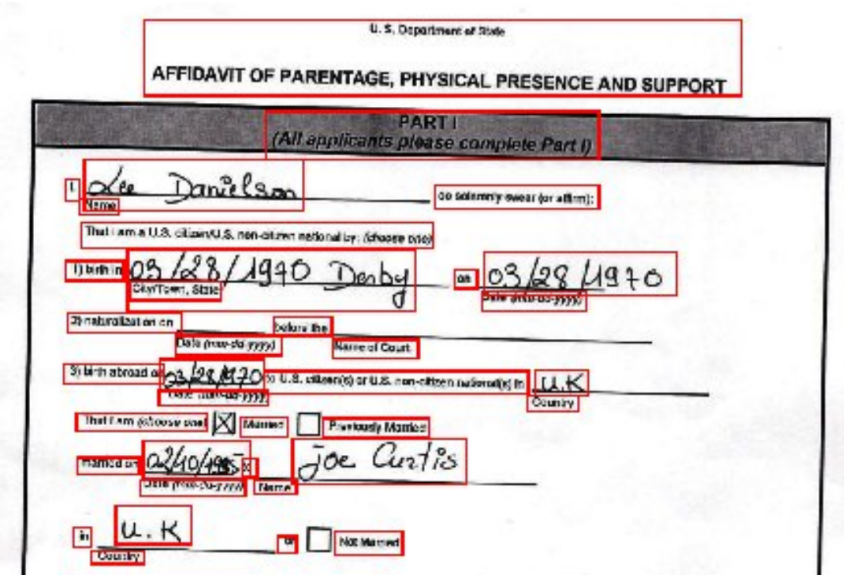
The Maurdor database

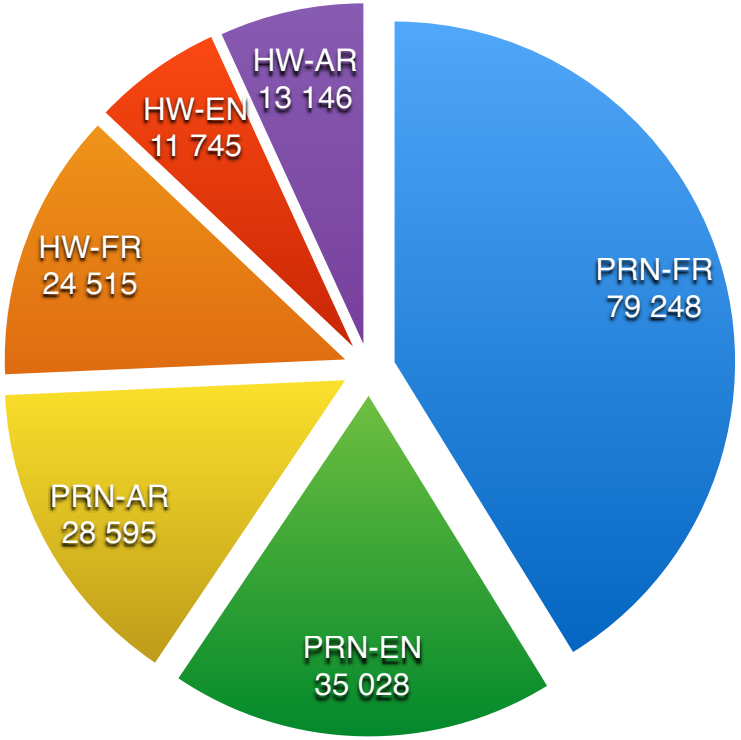


8,774 pages fully annotated : layout, images, graphics, languages, textual content, logical order.

The Maurdor challenge

Text recognition given the position, type (printed/handwritten) and language

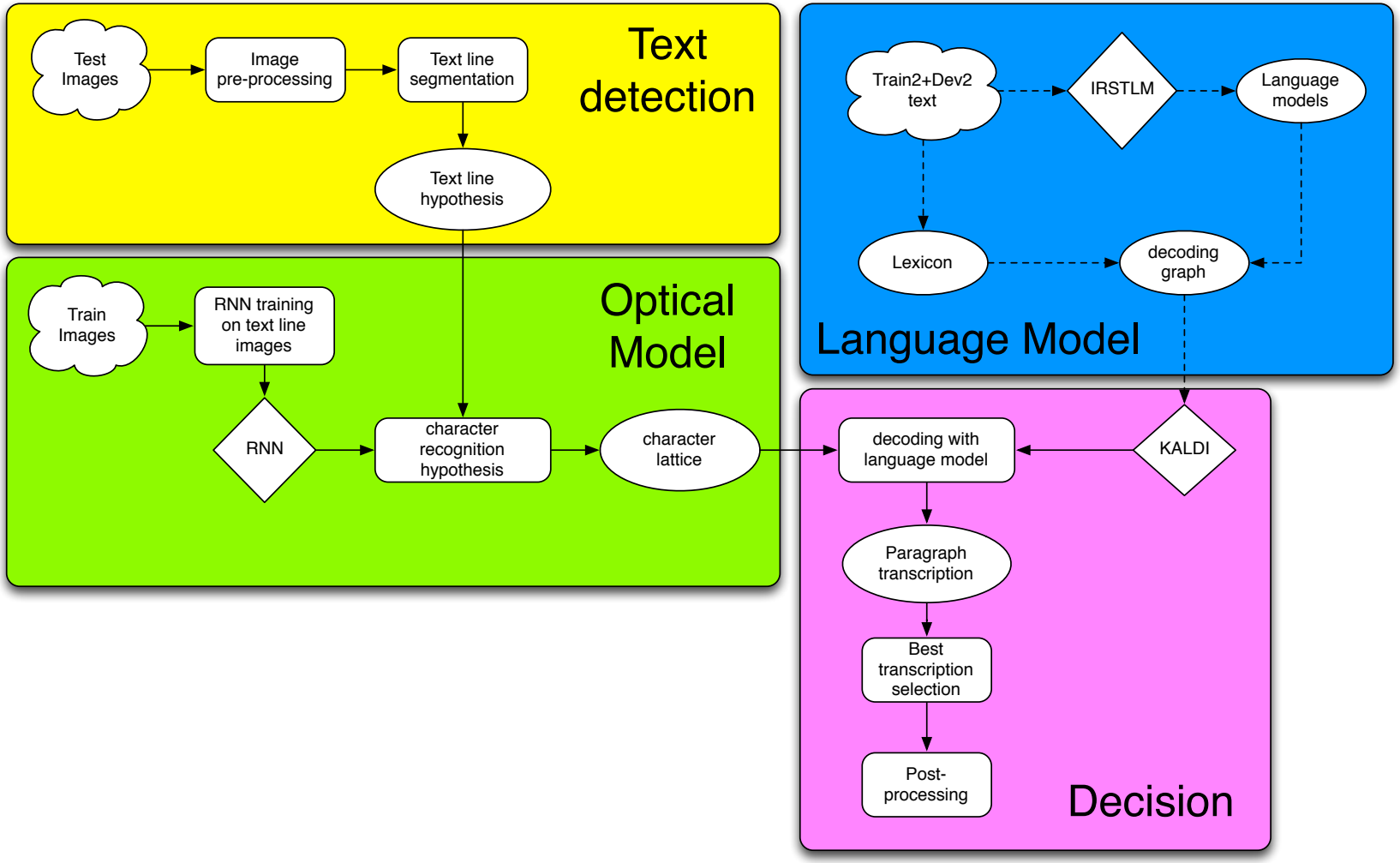




Text zone statistics per language and type

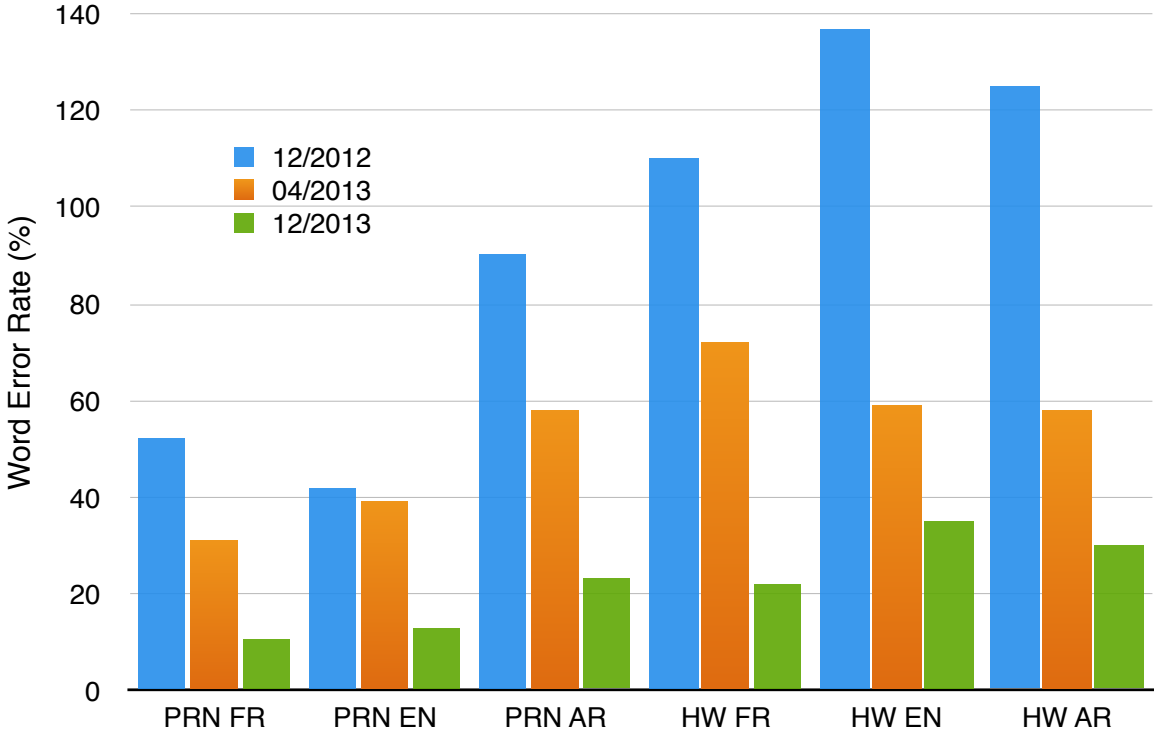
The A2iA system for this challenge

Same system for Printed/Handwritten, French/English/Arabic :



The A2iA system for this challenge

Performance improvement



Error rate divided by 3 to 5

Keys of success

Text detection

Consider several line detection hypothesis

Optical Model

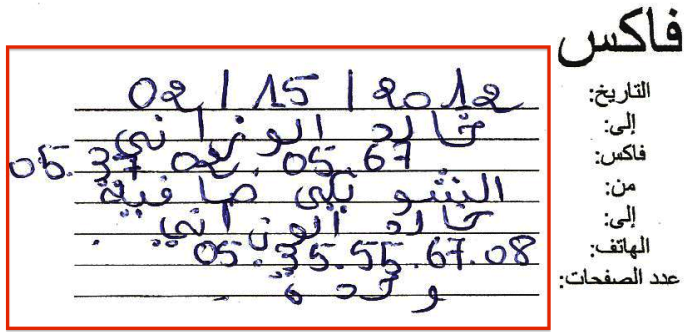
Train using curriculum

Model word and character sequences

Language Model

Text line detection

Difficult to find the text line, even if the text bloc is given

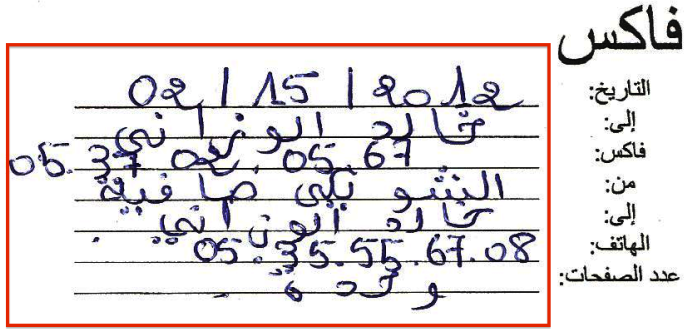


Text line detection

Difficult to find the text line, even if the text bloc is given

Generate hypothesis :

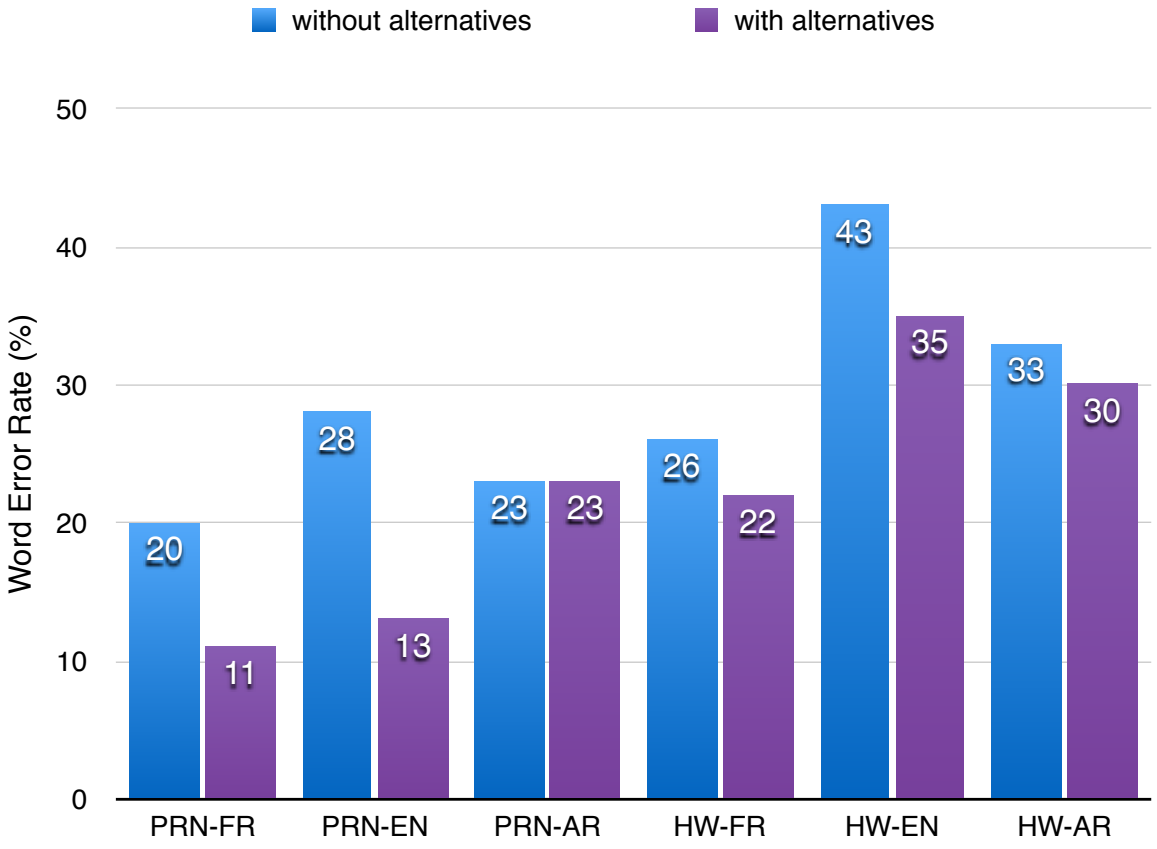
- use 2 different line detectors
- try with and without deskew
- try line compression and stretching
- normalize or not the line height
- re-order the text line
- try not to detect the line



Recognize all the hypothesis, keep the best

Text line detection

Results

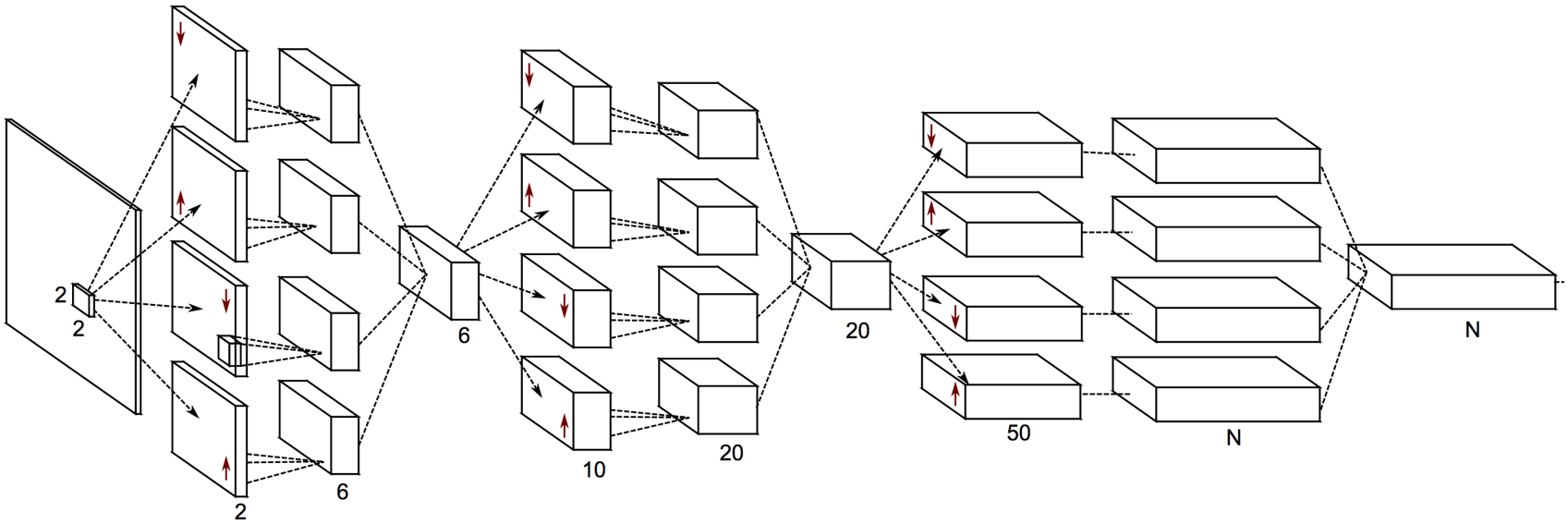


up to 50% word error rate reduction

Optical model : LSTM neural network

The RNN predicts hypothesis of characters :

- Latin letters with accent, upper and lower cases
- 120 shapes of Arabic forms
- punctuation, inter-word space and *not_a_character*



Optical model : LSTM neural network

Training

Curriculum training : start training with with easy samples and gradually increase the difficulty

- simple database at word level (Rimes, IAM, OpenHaRT)
- simple database at line level
- difficult database at line level (Maurdor)

Optical model

Improvements

RNN training set	# of training lines	Word error rate
Single lines	7310	54.7%
First step of automatic location	10570	43.8%
Second step of automatic location	10925	35.2%
Total number of lines (without location)	11608	-

40% word error rate reduction for handwritten English

Language models

Prepare the data :

- explicitly model the inter-word space
- separate punctuation signs and digit sequences (reduce the vocabulary)
- clean the textual data from unknown characters

Language models

Prepare the data :

- explicitly model the inter-word space
- separate punctuation signs and digit sequences (reduce the vocabulary)
- clean the textual data from unknown characters

Train the language model :

- Out-of-vocabulary words modelling with character n-gram
- In-vocabulary words sequence modelling with word n-gram
- for Arabic, use part-of-Arabic word decomposition

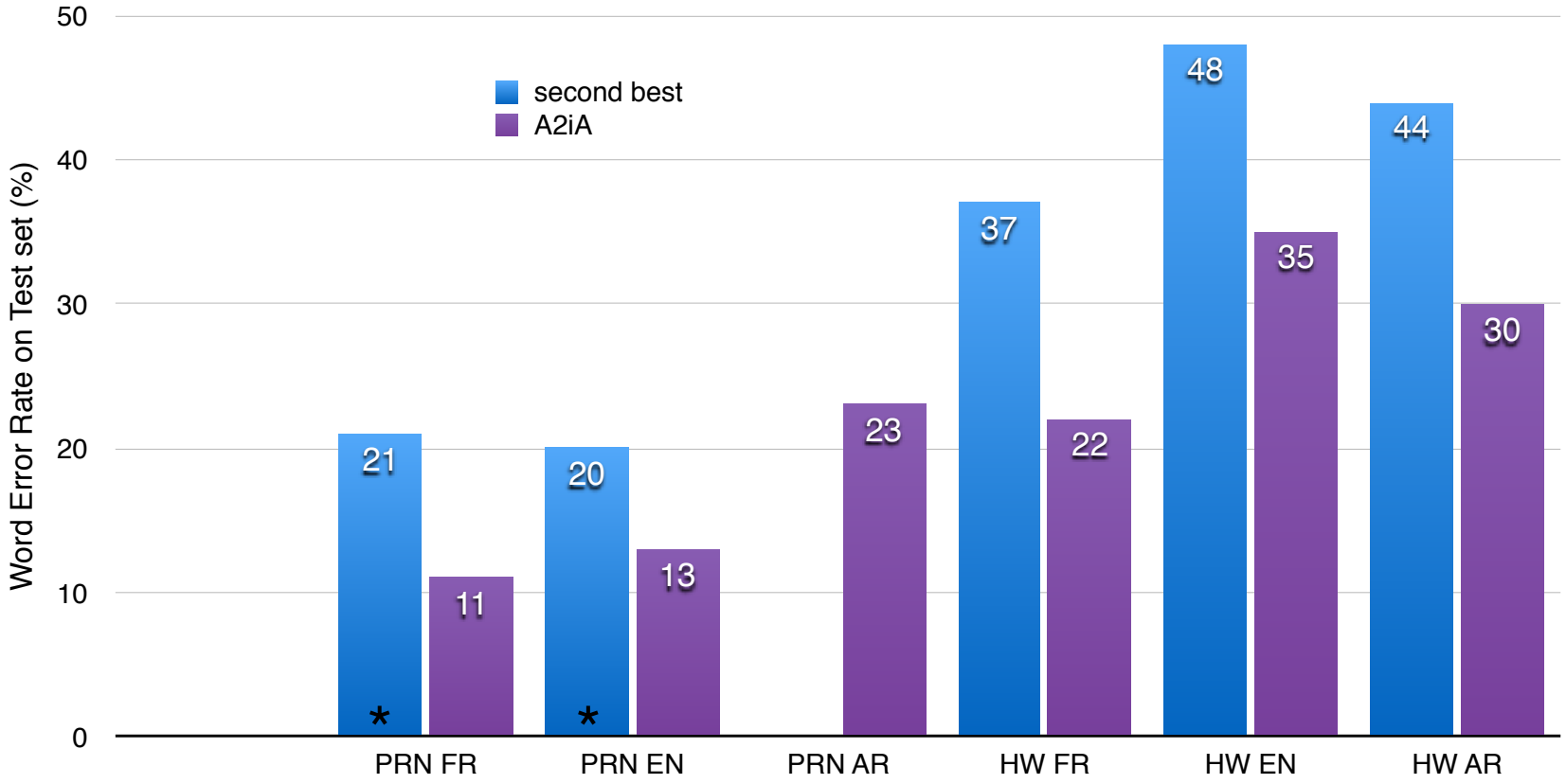
Language models

Language models evaluation :

Language	Type	PPL	%OOV	%Hit-Ratio		
				3-gram	2-gram	1-gram
English	PRN	111	3.9	37.1	41.8	21.1
	HWR	66	4.5	49.7	35.1	15.2
French	PRN	48	3.6	54.7	31.9	13.4
	HWR	73	3.9	48.6	36.6	14.7
Arabic	PRN	146	11.2	31.8	33.9	34.3
	HWR	134	8.4	29.5	44.9	25.6

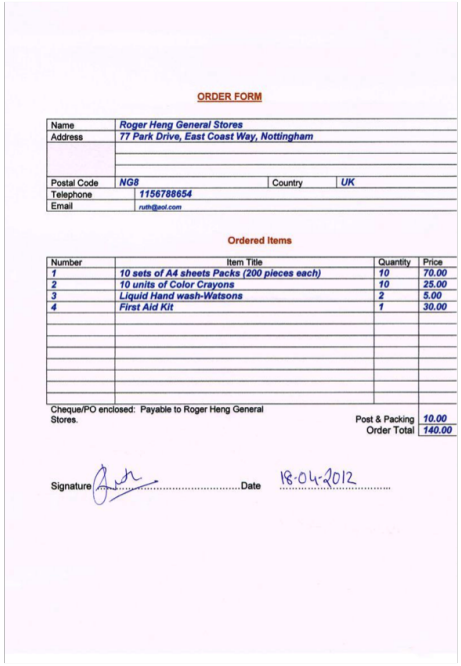
Maurdor evaluation

Maurdor second evaluation results :



Now ?

- Layout Analysis is the weak link
- Need for document understanding



ORDER FORM

Name	Roger Heng General Stores		
Address	77 Park Drive, East Coast Way, Nottingham		
Postal Code	NG8	Country	UK
Telephone	1156788654		
Email	ruth@aol.com		

Ordered Items

Number	Item Title	Quantity	Price
1	10 sets of A4 sheets Packs (200 pieces each)	10	70.00
2	10 units of Color Crayons	10	25.00
3	Liquid Hand wash-Watsons	2	5.00
4	First Aid Kit	1	30.00

Cheque/PO enclosed: Payable to Roger Heng General Stores.

Post & Packing 10.00
Order Total 140.00

Signature ..... Date 18-04-2012

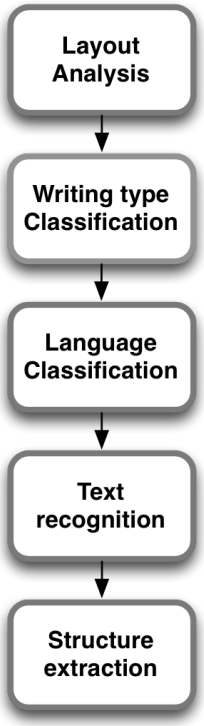
Name	Roger Heng General Stores
Address	77 Park Drive, East Coast Way, Nottingham

Quantity	Price
10	70.00
10	25.00
2	5.00
1	30.00

Post & Packing	10.00
Order Total	140.00

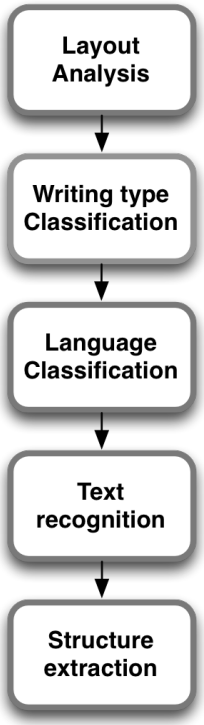
.Date 18-04-2012

The future of document processing ?

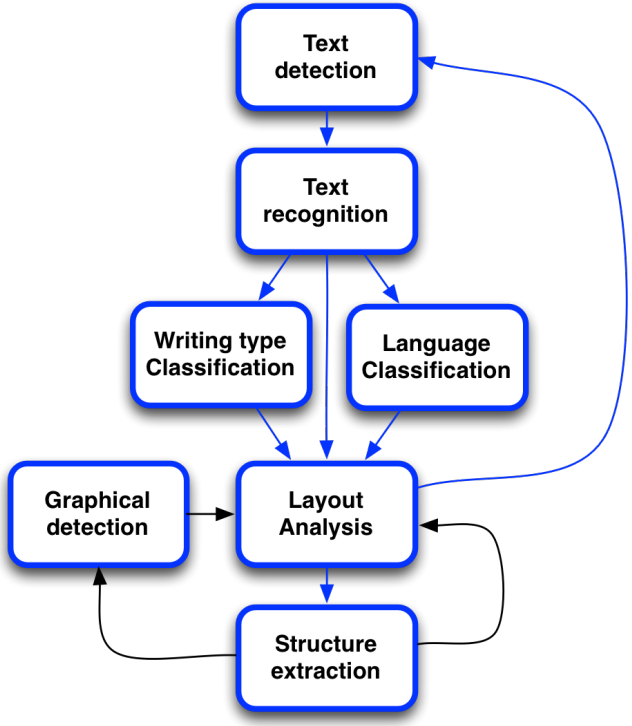


Initial Maurdor workflow

The future of document processing ?



Initial Maurdor workflow



Non linear workflow

Thank you !

www.a2ialab.com