Financé par

DGA

MAURDor

CASSIDIAN

The First MAURDOR Campaign

Olivier Galibert, Juliette Kahn, Ilya Oparin
firstname.secondname@lne.fr

MESURES & RÉFÉRENCES
Clés de la COMPÉTITIVITÉ et d'un MONDE PLUS SÛR

Laboratoire national de métrologie et d'essais

Evaluation Context

Financé par

More and more different types of documents to handle

There is a need for a complete chain to process scanned documents

- Detection of different zones in a document
- Identification of text type (handwritten/printed)
- Language identification
- Optical character recognition
- Revealing logical structure of documents

Retrieval information

MAURDOR Evaluation Campaign goals

Financé par

Evaluate existing tools for automatic processing of digitized documents

- Invoices, Bills, Fax headers, Forms, Letters
- Handwritten, Printed
- Mutilingual data : French, Arabic and English

Define state-of-the-art by evaluating performance of different systems

Campaign 2013: the first of its kind

Data

Created and annotated by ELDA (GEDI Format)

- C1 : Blank or completed forms (12%)
- C2 : Printed, but also manually annotated business documents (40%)
- C3 : Private handwritten correspondence, sometimes with printed letterheads (25%)
- C4 : Printed, but also manually annotated business correspondence (20%)
- C5 : Other documents such as newspaper articles or blueprints, etc. (3%)

Data distribution in train, dev and test sets made by LNE

- Homogeneous between train, dev and test
- Criteria of data partition
  - Language
  - Categories
  - Number of zones
  - Number of text zones
  - Number of words

1st Campaign : Corpus

		Train	Dev	Test
Number of documents		3,000	1,000	1,000
Partition in catégories (number of documents)	C1	13.3%	13.4%	14.7%
	C2	43.8%	41.7%	42.5%
	C3	20.8%	20.4%	21.2%
	C4	17.5%	19.9%	16.9%
	C5	4.5%	4.6%	4.7%
Partition regarding writing type	Printed	74.4%	75.4%	74.9%
	Handwritten	25.6%	24.7%	25.1%
Partition in languages (zones)	English	26.2%	24.7%	26.9%
	Arabic	22.0%	20.4%	23.1%
	French	51.8 %	54.8%	49.8%
	Other	42	17	13
Number of tokens		300,000	100,000	100,000

Regions repartition

Types	Subtypes	Train	Dev	Test
Text Region	-	74 955 (65.5%)	25 985 (66.4%)	25 676 (65.6%)
Graphic Region	All	25 229 (22.0%)	8 789 (22.5%)	8 725 (22.3%)
	Logos	1 245	458	390
	Stamp	16	6	6
	Signature	1 830	659	588
	Form fields	5 782	1 867	1 928
	Underlined form field	11 293	3 690	3 944
	Other form field	51	0	11
	drawing	1 436	678	401
	other	3 576	1 431	1 457
Picture region	-	588 (0.5%)	353 (0.9%)	139 (0.4%)
Drawing region	-	3 726 (3.3%)	1 431 (3.7%)	1 078 (2.8%)
Table region	-	830 (0.7%)	333 (0.9%)	253 (0.6%)
Linedrawing region	-	1 294 (1.1%)	464 (1.2%)	425 (1.1%)
Noise	-	7 873 (6.9%)	1 964 (5.0%)	2 853 (7.3%)

Tasks to Be Evaluated

Six tasks on the document level:

Module 1: Zone detection and classification

Module 2: Identification of writing type (handwritten/printed)

Module 3: Language identification (English, French, Arabic)

Module 4: Optical character recognition

Module 5: Extraction of the logical structure

End-to-end: Chain of modules from 1 to 5

Participants

Financé par

6 participants, 27 systems

Tasks	Number	Participants
Module 1	4	A2iA, IRISA, Jouve, LITIS
Module 2	2	IRISA, LITIS
Module 3	3	LIP 6, IRISA
Module 4	5	Jouve, RWTH, IRISA, LITIS, LIP 6
Module 5	3	Jouve, IRISA, LITIS
End to end	1	LITIS

Module 1: Task

Partitioning document images into distinct and homogeneous graphical areas

- Area delimitation using closed polygonal-shaped outlines
  - Zones may overlap
  - Page orientation

Zone classification

- Text zone
- Edge line
- Hand line drawing
- Damaged area
- Table
- Image
- Graphic zone (logo, graph, seal, signature, form field box, underlined field box,drawing, photo, other…)

ZoneMap

- Take in account split and merge situation
- Implementation details on the evaluation plan

Jaccard

$$J_i = \frac{H_i \cap R_i}{H_i \cup R_i}$$

Module 1: Results

Financé par

Zone types treated by the participants

	Syst.1	Syst.2	Syst. 3	Syst.4
	5	4	3	5
Types handled	• Text • Form field • Underlined form field • Table • Line drawing	• Text • Drawing • Other graphic zones • Table	• Text • Table • Line drawing	• Text • Logo • Signature • Table • Damaged area
% of potentially covered surface	41.8%	87.6%	40.4%	42.6%

Module 1: Results

Global results

System	Run	ZoneMap (%)	Jaccard
Syst.1	1	107.1	0.150
	2	90.7	0.169
	3	91.5	0.162
	4	91.6	0.162
Syst. 2	1	**57.3**	0.190
Syst. 3	1	76.0	0.315
	2	72.8	**0.382**
Syst. 4	1	62.5	0.287
	2	62.3	0.286

Module 1: Results by Document Category

ZoneMap

System	Run	C1	C2	C3	C4	C5
Syst 1	1	104.0	106.9	110.5	104.7	118.3
	2	87.5	91.8	95.1	82.4	87.8
	3	87.3	92.9	98.0	82.8	84.8
	4	88.6	93.0	94.8	82.2	89.3
Syst 2	1	**51.1**	**60.3**	**54.5**	**37.6**	70.9
Syst 3	1	76.1	79.2	64.4	51.4	84.2
	2	64.7	77.4	64.0	50.4	89.8
Syst 4	1	60.6	66.5	53.8	40.5	59.9
	2	62.3	66.4	51.6	37.8	**58.0**

Jaccard

System	Run	C1	C2	C3	C4	C5
Syst 1	1	0.184	0.125	0.105	0.324	0.253
	2	0.199	0.142	0.118	0.375	0.339
	3	0.198	0.133	0.114	0.245	0.366
	4	0.196	0.135	0.111	0.367	0.334
Syst 2	1	0.250	0.152	0.210	0.501	0.162
Syst 3	1	0.288	0.302	0.331	0.519	0.421
	2	**0.503**	**0.348**	**0.340**	0.544	0.313
Syst 4	1	0.331	0.237	0.292	0.586	**0.487**
	2	0.317	0.237	0.300	**0.621**	0.485

Financé par

Identification of writing type

- Handwritten
- Printed

$$P_{WritingType} = \frac{\text{number of correctly classified text zones}}{\text{number of text zones}}$$

$$P_{WritingTypeNP} = \frac{\text{number of not classified text zones}}{\text{number of text zones}}$$

Module 2: Results

Global

System	Précision (%)	Silence (%)
Syst.1	**90.4**	6.6
Syst. 2	89.9	0.0

By writing type

System	Printed			Handwritten		
	Precision (%)	Recall (%)	F-measure	Precision (%)	Recall (%)	F-measure
Syst.1	92.3	**95.5**	**93.9**	**82.6**	73.0	77.5
Syst. 2	**93.9**	92.5	93.2	78.5	**82.1**	**80.3**

By language

System	Precision (%)				Silence (%)			
	ENG	ARB	FRA	Other	ENG	ARB	FRA	Other
Syst.1	**86.3**	89.3	**93.0**	90.2	8.1	8.6	4.8	1.9
Syst. 2	85.7	**93.0**	90.6	**96.2**	0.0	0.0	0.0	0.0

By document category

	Precision (%)					Silence (%)				
	C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
Syst.1	92.4	**88.6**	**92.1**	**91.0**	**86.4**	8.2	5.5	5.0	4.4	4.5
Syst. 2	**92.9**	88.3	90.5	89.2	81.6	0.0	0.0	0.0	0.0	0.0

Financé par

Language identification

- French
- English
- Arabic
- Other

$$P_{LanguageBlock} = \frac{\text{number of zones with correctly identified language}}{\text{total number of text zones}}$$

$$P_{LanguageBlockNP} = \frac{\text{number of zones with no language identified}}{\text{total number of text zones}}$$

Module 3: Results

Global

System	Run	Précision (%)	Silence (%)
Syst 1	1	38.9	0.0
	2	33.1	1.1
Syst 2	1	**63.8**	0.0
	2	58.1	0.0

By writing type

System	Run	Précision (%)		Silence (%)	
		Printed	Handwritten	Printed	Handwritten
Syst 1	1	35.4	49.2	0.0	0.0
	2	30.5	39.5	0.9	1.5
Syst 2	1	**64.5**	**61.7**	0.0	0.0
	2	61.0	49.5	0.0	0.0

By document category

System	Run	Precision (%)					Silence (%)				
		C1	C2	C3	C4	C5	C1	C2	C3	C4	C5
Syst 1	1	42.7	34.4	49.8	47.3	27.2	0.0	0.0	0.0	0.0	0.0
	2	38.4	27.5	44.6	39.4	19.4	0.9	1.3	0.1	0.6	2.0
Syst 2	1	**71.7**	**53.2**	**62.3**	**60.5**	**65.2**	0.0	0.0	0.0	0.0	0.0
	2	65.1	49.5	58.2	55.8	55.9	0.0	0.0	0.0	0.0	0.0

Module 3: Results by Language

Financé par

System	Run	Precision (%)			Recall (%)			F-measure		
		ENG	ARB	FRA	ENG	ARB	FRA	ENG	ARB	FRA
Syst 1	1	**41.7**	28.7	58.8	**27.5**	69.7	30.9	**33.1**	40.7	40.5
	2	29.8	25.2	54.1	7.9	72.2	28.5	12.4	37.4	37.3
Syst 2	1	-	**75.5**	**60.4**	0.0	**73.6**	**94.0**	-	**74.5**	**73.5**
	2	-	55.7	59.0	0.0	65.2	86.5	-	60.1	70.2

Financé par

Task: OCR (transcription)

$$WER = \frac{N_i + N_d + N_s + \chi N_{np}}{\text{number of words in the reference}}$$

$$WER = \frac{N_i + N_d + N_s + \chi N_{np}}{\text{number of characters in the reference}}$$

Module 4: Results by Writing Type and Language

Financé par

CER

System	Run	Printed					Handwritten				
		ARB	FRA	ENG	latin	All	ARB	FRA	ENG	latin	All
Syst.1	1	**39.8**	17.1	25.4	20.6	22.7	**31.9**	41.4	32.8	39.3	**37.2**
Syst 2	1	54.7	**11.3**	**13.1**	**12.1**	**16.7**	-	75.4	84.5	77.6	83.9
Syst 3	1	98.1	79.3	88.0	82.9	84.6	112.6	101.7	124.4	107.2	108.7
Syst 4	1	65.4	27.7	29.2	28.4	32.5	82.3	67.2	85.3	71.6	74.6
Syst 5	1	-	-	-	-	-	-	24.5	22.2	24.0	45.3
	2	-	-	-	-	-	-	24.2	22.3	23.8	45.1
	3	-	-	-	-	-	-	22.2	21.8	22.1	43.9
	4	-	-	-	-	-	-	**21.3**	**20.3**	**21.1**	43.2

Module 4: Results by Writing Type and Language

WER

System	Run	Typographique					Handwritting				
		ARB	FRA	ENG	latin	all	ARB	FRA	ENG	latin	all
Syst.1	1	**58.3**	31.0	39.2	34.4	37.1	**58.0**	71.6	58.7	68.1	65.4
Syst 2	1	91.3	**21.0**	**20.9**	**20.9**	**28.9**	-	98.0	102.4	99.2	99.4
Syst 3	1	161.3	141.6	160.4	149.4	150.7	149.5	173.0	201.0	180.6	172.2
Syst 4	1	123.9	64.0	66.8	65.1	70.5	101.0	97.2	118.0	102.8	102.3
Syst 5	1	-	-	-	-	-	-	40.6	41.6	40.9	56.7
	2	-	-	-	-	-	-	39.8	41.8	40.3	56.3
	3	-	-	-	-	-	-	37.7	40.6	38.5	55.0
	4	-	-	-	-	-	-	**35.9**	**38.1**	**36.5**	**53.5**

Module 5: Task and Metrics

Task: Extraction of the logical structure

Three criteria are defined:

- A semantic subtype (for example header, text body, etc.) +1 point
- Zone that precedes the one in question +1point
- Set *E* of zones present in the same group as the one in question (F-measure)

Each criterion gives a per-document score

Normalization is done relative to S_0 = score when there is no answer

$$S_b \leq S_0, S = 100\frac{S_b - S_0}{S_0} \qquad S = 100\frac{S_b - S_0}{1 - S_0}$$

Module 5: Results

Global

System	Type	Ordre	Group
Syst.1	**59**	**22**	26
Syst. 2	42	17	**37**
Syst. 3	11	2	26

By document category

		Type	Ordre	Groupe
C1	Syst.1	**38**	**46**	58
	Syst. 2	0	33	**73**
	Syst. 3	-1	5	0
C2	Syst.1	**48**	1	0
	Syst. 2	34	**8**	**5**
	Syst. 3	2	1	0
C3	Syst.1	**68**	0	28
	Syst. 2	54	0	2
	Syst. 3	16	0	**67**
C4	Syst.1	**73**	**51**	33
	Syst. 2	52	14	**65**
	Syst. 3	23	0	10
C5	Syst.1	**4**	0	0
	Syst. 2	-1	**100**	0
	Syst. 3	1	**100**	0

Evaluation End-to-end: Task and Metrics

Similar to evaluation of search engines

List of quasi-words is constructed

- Wordlist sorted by frequency, second third is retained
- Information about zone types from Module 1 and about the logical function from Module 5 are added
- For each document the list of important words is constructed

Two metrics are used

- Standard cosine distance used in IR
- Utility metric based on word types
  - Found word scores +1
  - Extra word scores -1
  - Missed word scores 0

Evaluation End-to-end: Results

Financé par

Global

System	Global		Text		Type		Fonction	
	Cosinus	Utility	Cosinus	Utility	Cosinus	Utility	Cosinus	Utility
Syst.1	+0.4595	+0.0923	+0.2952	-0.0925	+0.8260	+0.3454	+0.2571	+0.0240

1st campaign results conclusions

Module 1

- ZoneMap and Jaccard metrics are additionnal
- No participants handles all the types of zones
- All participants handles text and table zones
- Text zones are the best detected
- The best results are obtained for the documents of category C4

Module 2

- Precision of the systems is around 90%
- Results are homogenous across document categories
- Printed writing is better recognized

1st campaign results conclusions

Module 3

- Best precision is around 60%

Module 4

Results depend on language, writing type and document category

Module 5

- All the systems add useful information
- The best results are obtained for document categories C3 and C4

Conclusions

Financé par

All the modules have been evaluated

- 6 participants, 27 systems

Generally, the best results are obtained for the documents of category C4

The same evaluation protocol will be kept for the second campaign

Next campaign

Second campaign starts in November 2013

- More data will be available
  - Train : 6000 documents
  - Dev : 1000 documents
  - Test : 1000 documents

- Same rules as for the 1st campaign

- Evaluation plan available on maurdor-campaign.org

Thanks for your attention

Juliette.kahn@lne.fr

www.maurdor-campaign.org

MESURES & RÉFÉRENCES
Clés de la COMPÉTITIVITÉ et d'un MONDE PLUS SÛR