EVALUATION OF ARTIFICIAL INTELLIGENCE SYSTEMS

GUILLAUME AVRIN

18/10/2019

LABORATOIRE
NATIONAL
DE MÉTROLOGIE
ET D'ESSAIS
LNE

Our positioning in AI

EVALUATION OF ARTIFICIAL INTELLIGENCE SYSTEMS

LNE, state-owned trusted third party for the evaluation of AI and robots

As a state-owned laboratory:

It is independent of any private interest (reinforced notion of trusted third party)

The sincerity of its evaluations is guaranteed

More than 10 years of experience on AI evaluation and more than 900 systems evaluated by a permanent team of doctors and engineers specialized in evaluation**.**

1. Assistance to public bodies

Organize evaluation campaigns

Develop evaluation methods and metrics

As a public research laboratory

Objective: measure technological progress and estimate investment impact to optimize public funding of research.

To developers

2. Technical assistance to companies

Development

Comparison (challenges)

Acceptance testing

To end users

Objective: provide our partners with reliable benchmarks and results to enable a pragmatic and well-reasoned decision-making.

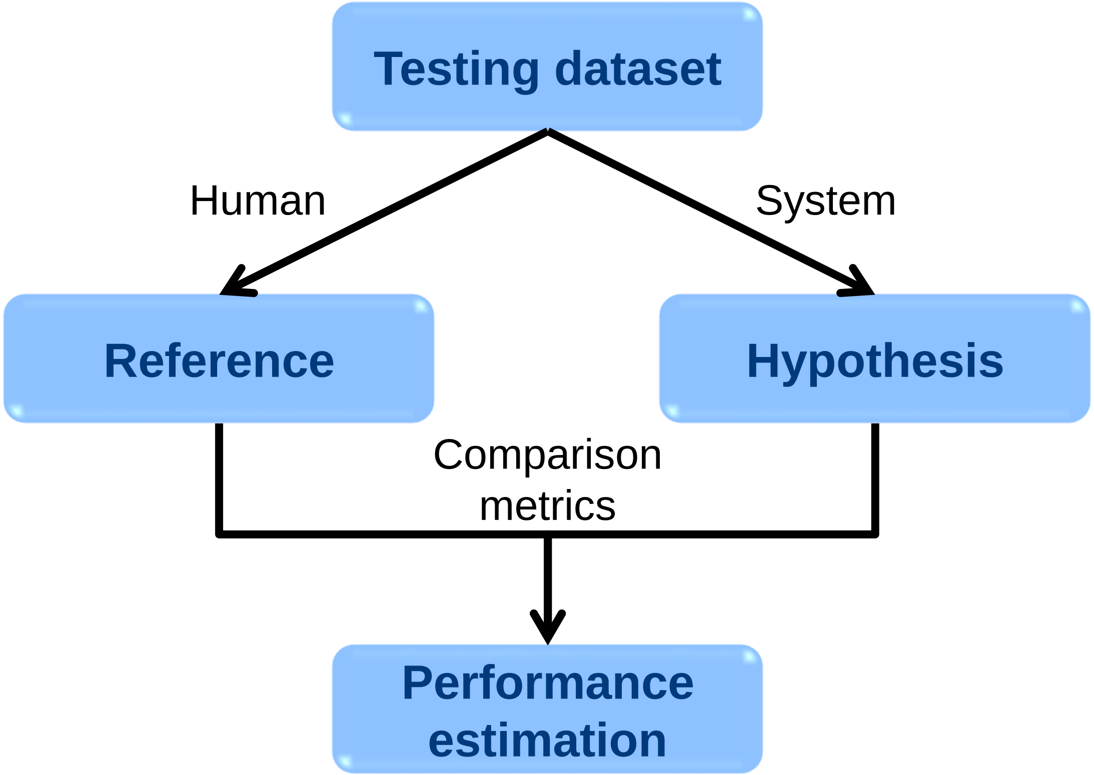
3. Participation to standardization activities

AFNOR AI, ISO AI, UNM 81, etc.

Objective: establish benchmarks to simplify contractual relations and encourage innovation while ensuring the protection of citizens and consumers.

The evaluation, step by step

AI EVALUATION PROCESS



Testing dataset

Human

System

Reference

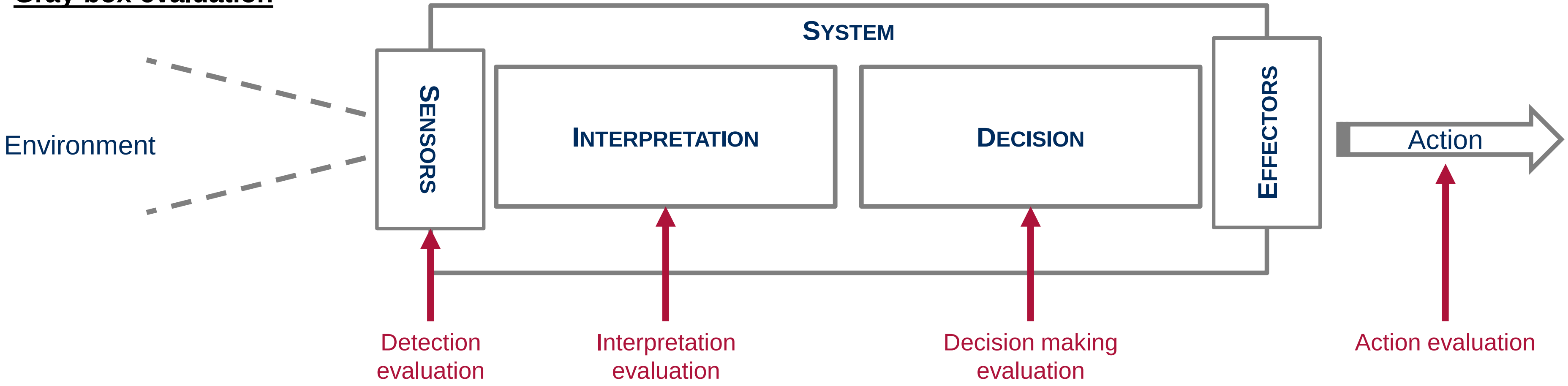
Hypothesis

Comparison metrics

Performance estimation

AI EVALUATION PROCESS

Gray-box evaluation



Black-box evaluation

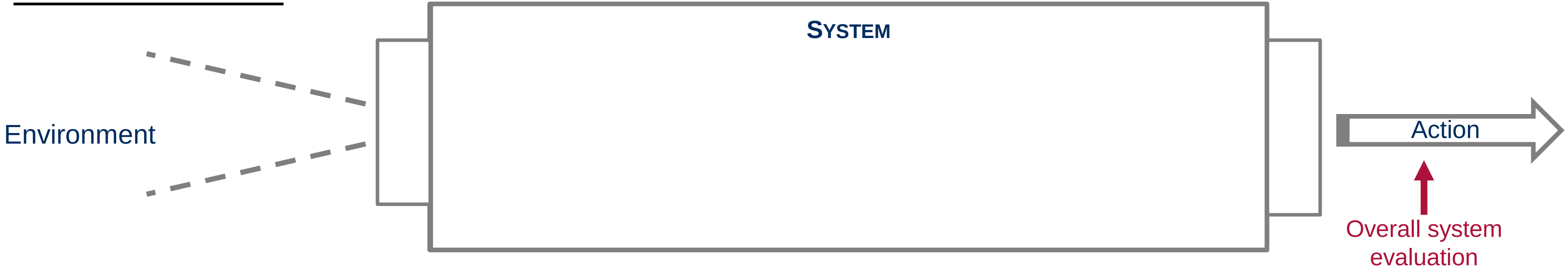
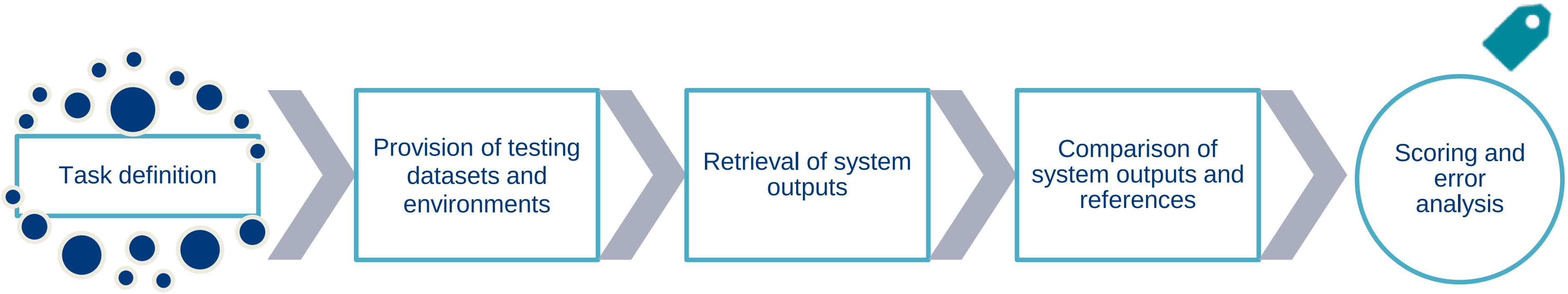


ILLUSTRATION OF THE STEPS OF AN EVALUATION



Task definition

Provision of testing datasets and environments

Retrieval of system outputs

Comparison of system outputs and references

Scoring and error analysis

EVALUATION : AN EXPERTISE IN ITS OWN RIGHT

Evaluation plan		Evaluation references		
1. Testing scenarios	2. Protocols, metrics	3. Testing environments	4. Data	5. References (ground truth)
Identification of technoscientific barriers to be removed	Identification of influencing factors	Development of adapted testing environments	Data selection: relevance, representativeness, quality	Development of annotation systems
Definition of participation terms and conditions	Definition of evaluation criteria and metrics	Control and measure of influencing factors	Development of tools for data management and sharing (server)	Data annotation or supervision of data annotation
Definition of the evaluation tasks	Interpretation of results	Ensure reproducibility of experiments	Development of tools for data collection	Qualification of annotations and annotators

LNE EVALUATION TOOLS

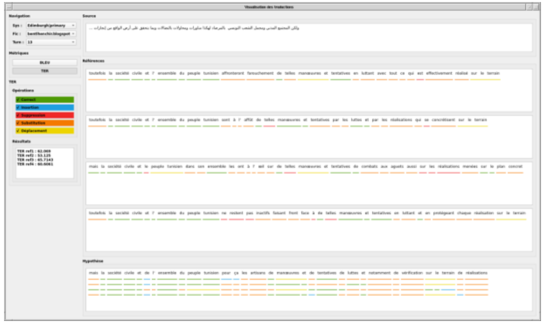
1. Testing scenarios

2. Protocols, metrics

3. Testing environments

4. Data

5. References (ground truth)



Visualisation des traductions

Navigation

Sys : Edinburgh/primary

Fic : benthbenchir.blogspot

Turn : 13

Métriques

BLEU

TER

TER

Opérations

✓ Correct

✓ Insertion

✓ Suppression

✓ Substitution

✓ Déplacement

Résultats

TER ref1 : 62.069
TER ref2 : 53.125
TER ref3 : 65.7143
TER ref4 : 60.6061

Source

... ولكن المجتمع المدني ويحمل الشعب التونسي بالمرصاد لهكذا مناورات ومحاولات بالنضالات وبما يتحقق على أرض الواقع من إنجازات

Références

toutefois la société civile et l' ensemble du peuple tunisien affronteront farouchement de telles manœuvres et tentatives en luttant avec tout ce qui est effectivement réalisé sur le terrain

toutefois la société civile et l' ensemble du peuple tunisien sont à l' affût de telles manœuvres et tentatives par les luttes et par les réalisations qui se concrétisent sur le terrain

mais la société civile et le peuple tunisien dans son ensemble les ont à l' œil sur de telles manœuvres et tentatives de combats aux aguets aussi sur les réalisations menées sur le plan concret

toutefois la société civile et l' ensemble du peuple tunisien ne restent pas inactifs faisant front face à de telles manœuvres et tentatives en luttant et en protégeant chaque réalisation sur le terrain

Hypothèse

mais la société civile et de l' ensemble du peuple tunisien pour ça les artisans de manœuvres et de tentatives de luttes et notamment de vérification sur le terrain de réalisations

Open-source Matics software suite to explore annotated data and evaluation results:

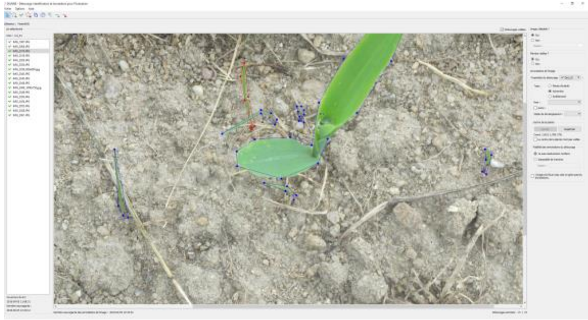
- Translation
- Diarization
- Transcription
- Speaker verification

And soon:

- OCR
- Image recognition

Evaluation of robots:

- laboratory testing (in LNE climatic chambers)
- virtual testing (simulation-based)



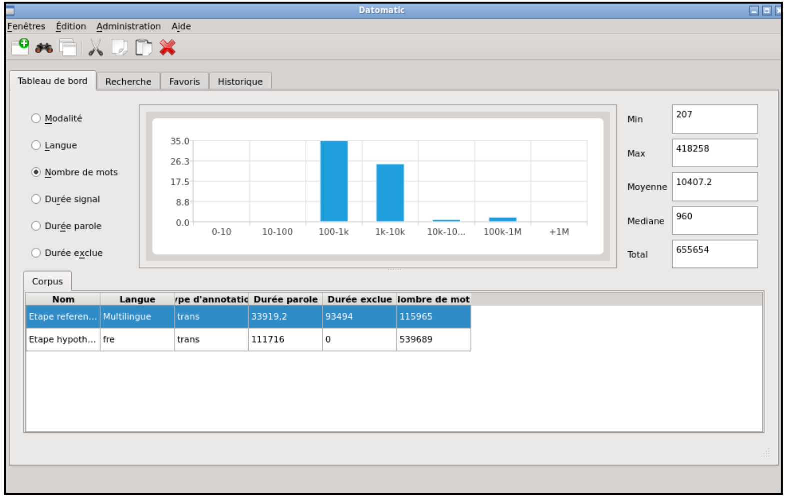
DIANNE software:

- annotation and automatic pre-annotation of crops and weed
- will be extended to other recognition tasks

OUR TOOLS

Matics software suite – Data visualisation and evaluation

Datomatic – Dataset preparation and visualisation
Evalomatic – Evaluation and visualisation

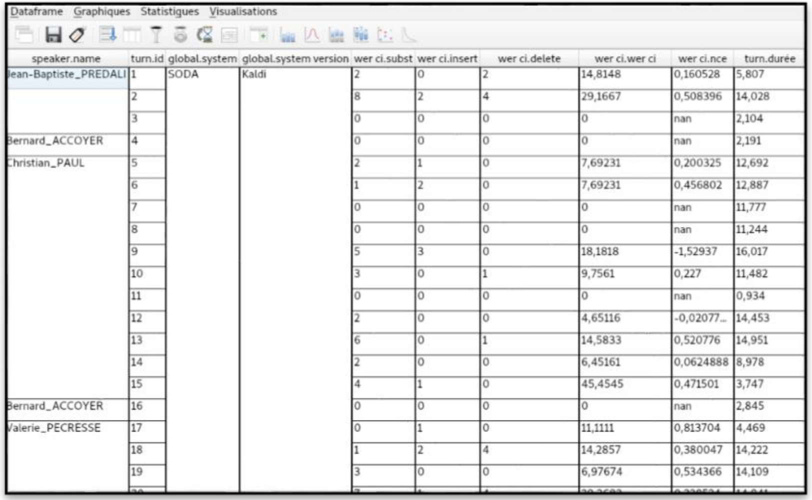


○ Modalité
○ Langue
● Nombre de mots
○ Durée signal
○ Durée parole
○ Durée exclue

Min 207
Max 418258
Moyenne 10407.2
Mediane 960
Total 655654

Nom	Langue	ype d'annotatic	Durée parole	Durée exclue	lombre de mot
Etape referen...	Multilingue	trans	33919,2	93494	115965
Etape hypoth...	fre	trans	111716	0	539689

Datomatic
Data visualisation
(transcription task)

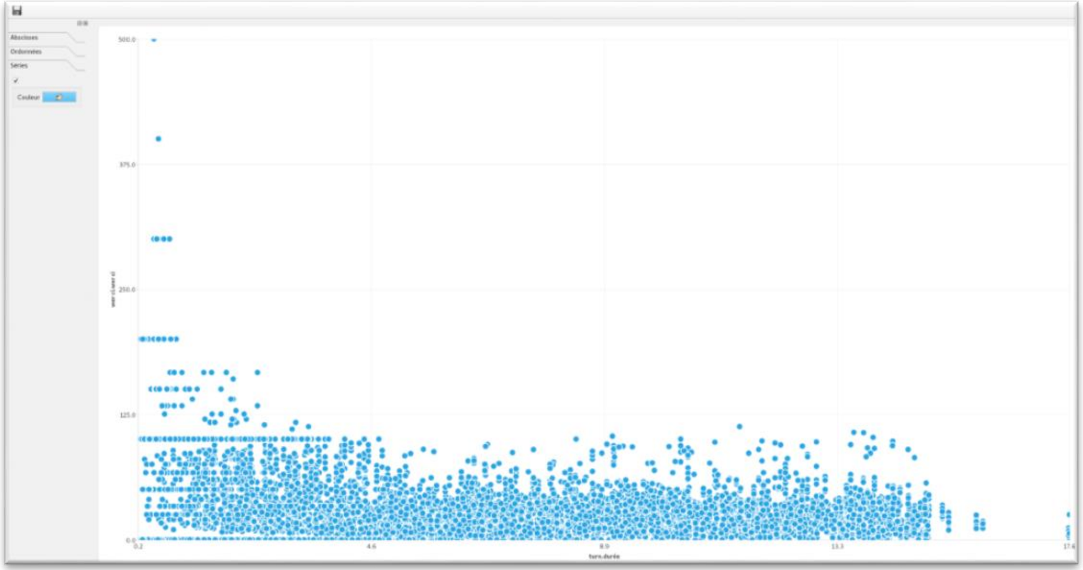


speaker.name	turn.id	global.system	global.system version	wer ci.subst	wer ci.insert	wer ci.delete	wer ci.wer ci	wer ci.nce	turn.durée
Jean-Baptiste_PREDALI	1	SODA	Kaldi	2	0	2	14,8148	0,160528	5,807
	2			8	2	4	29,1667	0,508396	14,028
	3			0	0	0	0	nan	2,104
Bernard_ACCOYER	4			0	0	0	0	nan	2,191
Christian_PAUL	5			2	1	0	7,69231	0,200325	12,692
	6			1	2	0	7,69231	0,456802	12,887
	7			0	0	0	0	nan	11,777
	8			0	0	0	0	nan	11,244
	9			5	3	0	18,1818	-1,52937	16,017
	10			3	0	1	9,7561	0,227	11,482
	11			0	0	0	0	nan	0,934
	12			2	0	0	4,65116	-0,02077...	14,453
	13			6	0	1	14,5833	0,520776	14,951
	14			2	0	0	6,45161	0,0624888	8,978
	15			4	1	0	45,4545	0,471501	3,747
Bernard_ACCOYER	16			0	0	0	0	nan	2,845
Valerie_PECRESSE	17			0	1	0	11,1111	0,813704	4,469
	18			1	2	4	14,2857	0,380047	14,222
	19			3	0	0	6,97674	0,534366	14,109

Evalomatic
Evaluation scores
(transcription task)

OUR TOOLS

Matics software suite

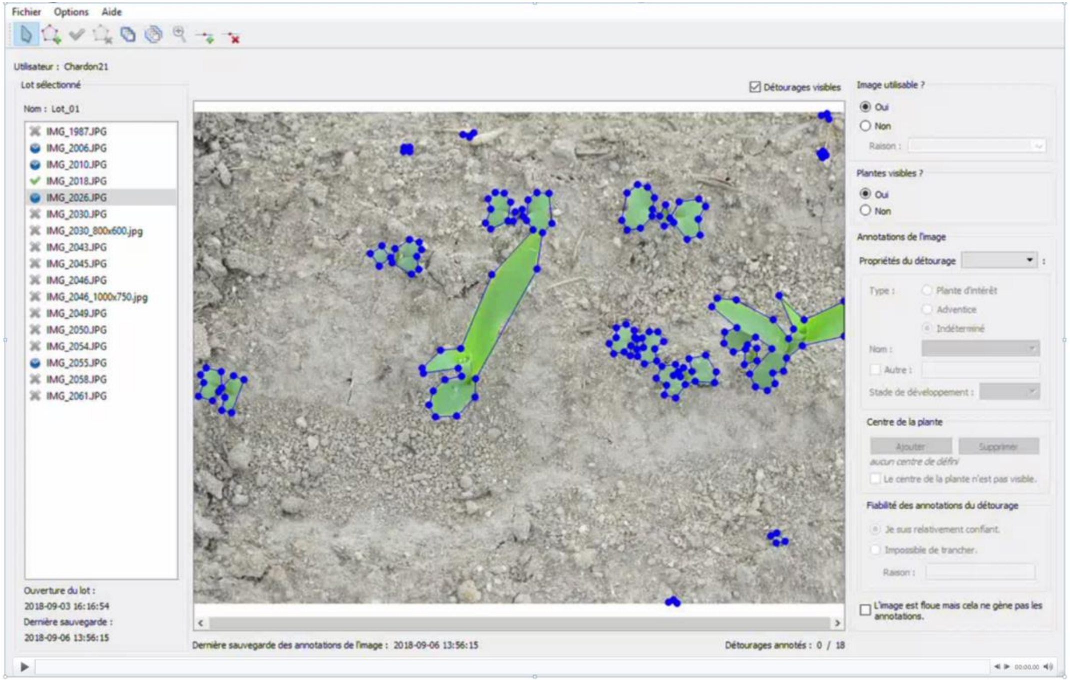


Evalomatic

Graphical visualisation

OUR TOOLS

DIANNE : Edge detection, identification and annotation for evaluation



Utilisateur : Chardon21

Lot sélectionné

Nom : Lot_01

- IMG_1987.JPG
- IMG_2006.JPG
- IMG_2010.JPG
- ✓ IMG_2018.JPG
- IMG_2026.JPG
- IMG_2030.JPG
- IMG_2030_800x600.jpg
- IMG_2043.JPG
- IMG_2045.JPG
- IMG_2046.JPG
- IMG_2046_1000x750.jpg
- IMG_2049.JPG
- IMG_2050.JPG
- IMG_2054.JPG
- IMG_2055.JPG
- IMG_2058.JPG
- IMG_2061.JPG

Ouverture du lot :
2018-09-03 16:16:54
Dernière sauvegarde :
2018-09-06 13:56:15

☑ Détourages visibles

Dernière sauvegarde des annotations de l'image : 2018-09-06 13:56:15

Détourages annotés : 0 / 18

Image utilisable ?
● Oui
○ Non
Raison :

Plantes visibles ?
● Oui
○ Non

Annotations de l'image

Propriétés du détourage :

Type : ○ Plante d'intérêt
○ Adventice
● Indéterminé

Nom :

□ Autre :

Stade de développement :

Centre de la plante

aucun centre de défini

□ Le centre de la plante n'est pas visible.

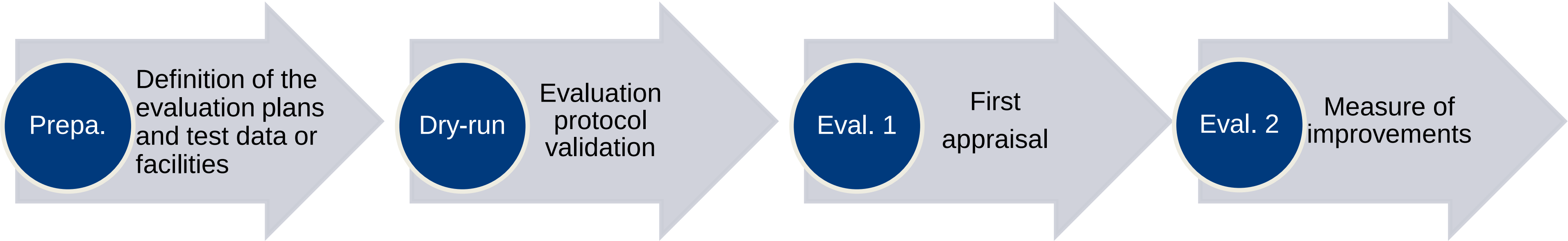
Fiabilité des annotations du détourage

● Je suis relativement confiant.
○ Impossible de trancher.
Raison :

□ L'image est floue mais cela ne gène pas les annotations.

00:00.00

CHALLENGE ORGANISATION



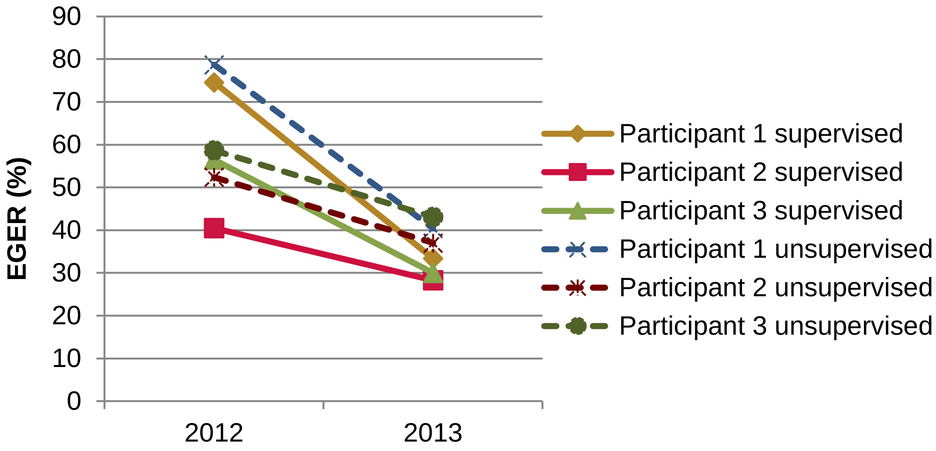
Prepa. – Definition of the evaluation plans and test data or facilities

Dry-run – Evaluation protocol validation

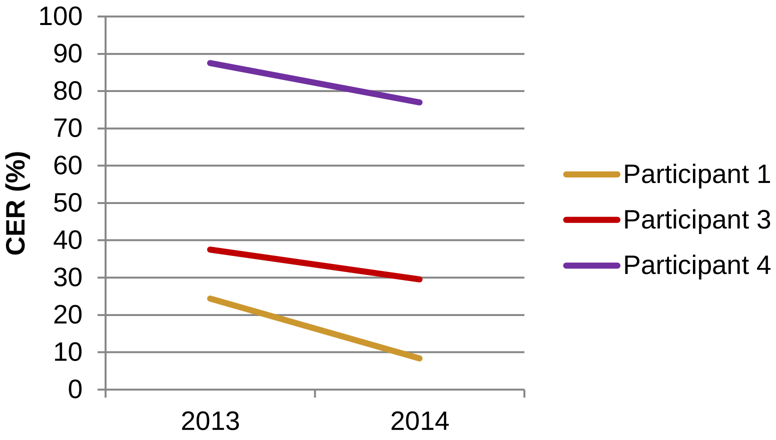
Eval. 1 – First appraisal

Eval. 2 – Measure of improvements

Evolution of error rates – person recognition
(REPERE campaign)



Evolution of error rates – optical character recognition
(MAURDOR campaign)

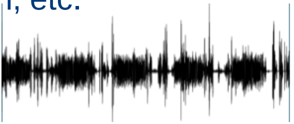


For which application areas?

EXPERTISE IN EVALUATION OF INFORMATION PROCESSING SYSTEMS

Speech

Transcription, keyword spotting, speaker comparison, named entities recognition, speaker tracking, translation, etc.



Text

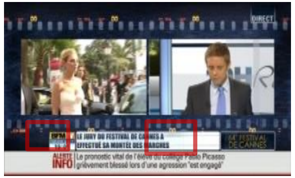
Topic detection, named entities recognition, information retrieval, translation, etc.

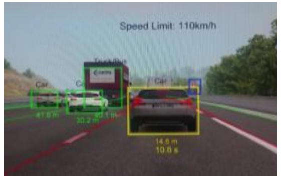
ترحيب ، يسعدنا أن نرحب بكم

Welcome, we are delighted to have you here

Image

Head tracking, optical character recognition, etc.


DIRECT
LE JURY DU FESTIVAL DE CANNES A EFFECTUÉ SA MONTÉE DES MARCHES
64E FESTIVAL DE CANNES
ALERTE INFO
Le pronostic vital de l'élève du collège Pablo Picasso grièvement blessé lors d'une agression "est engagé"



Multimedia

Person tracking, document classification, etc.

- Challenges (Quaero, Repere, etc.)
- Benchmarking (INC)
- Qualification (Allies)
- Certification (Voxcrim)

EVALUATION OF ROBOTS

- **Smart mobility** Simulation for autonomous vehicle safety
- **Agri-food** Risk analysis, scientific monitoring and community structuring, organization of a challenge in agricultural robotics
- **Service** Development of evaluation tools
- **Public-Private partnership** Study of the influence of climatic conditions on the performance of AI systems, assessment of AI and cybersecurity of smart medical devices

Simulation of the autonomous vehicle

HRP2 robot (Franco-Japanese humanoid robot) evaluated in climatic chambers at LNE

Our orientations

OUR ORIENTATIONS

Metrology: develop standards and protocols for the evaluation of AI

Evaluation: set up an AI assessment and testing centres

Certification: promote the certification of AI

METROLOGY OF AI

Definition of standards: reference testing datasets and environments, metrics, etc.

Definition of evaluation protocols: testing scenarii, evaluation tasks, methods for calculating the measurement uncertainty, etc.

For performance evaluation

- Accuracy, precision, trueness, fidelity, error rate, sensitivity, specificity, etc.
- Robustness, resilience and operating range
- Datasets qualification (representativeness)
- Other performance requirements (speed, efficiency, ergonomics, etc.)

To promote acceptability

- Regulation (transparency, non-discrimination)
- Explainability, intelligibility, predictability, readable behaviour
- Security (controllable, auditable)

EXPLAINABILITY

A tool to facilitate verifications and make them more reliable

- Solving the “black box” problem?
- To estimate the operating domain, better identify rare (but critical) phenomena, etc.

Towards the evaluation of explainability

- Measuring performance
  - Characterise explainability (according to context, requirements, user profile, etc.)
  - Define objective metrics
- Development of standards
  - Type of information to be extracted, reference values, etc.

CONCLUSION

What LNE offers:

- A unique know-how in the organization of evaluation campaigns for AI systems (design of the evaluation plan, organization of evaluation meetings, management of the associated events)
    - to set up a rigorous metrological approach (repeatable performance measurements, reproducible experiments, qualified test databases, identified and controlled influence factors, limited biases)
    - to maximize the impact of evaluations
- Evaluation tools
    - Suite Matics software suite
    - annotation tools
    - real or simulated test environments, etc.
- A status:
    - trusted third party (LNE does not develop AI systems)
    - independent evaluator (LNE is public, it is independent of any private interest)

LNE is interested in:

- collaborating with other **NMI** to bring metrology expertise to the field of AI evaluation.
- participating in projects aimed at demonstrating the performance and functionality of AI technologies
- setting up challenges, evaluation campaigns, competitions, especially in AI and robotics

COLLABORATIVE TOPICS WITH LNE

Performance evaluation: accuracy, precision, trueness, robustness, resilience

- at the level of the overall system (autonomous cars, surgical robots, etc.),
- at the level of the detection modules (obstacle detection, face recognition, etc.),
- at the level of decision-making modules (hazard management, etc.),
- At the level of action modules (autonomous navigation, etc.).

Explainability evaluation: how to relate the decision taken to the known data and characteristics of the situation?

Human-machine interaction evaluation: how to measure the quality of an interaction (during a close cooperation between an intelligent personnel assistant and a pilot, for example).

Thank you for your attention