

Hybridation des approches symboliques et d'apprentissage profond pour la reconnaissance des entités dans les signatures de mail

Hugo Thomas, Nicolas Fouqué, Duc Hau Nguyen, Victor Klötzer



Remerciements aux organisateurs



Equipe LINKMEDIA



Rennes, Bretagne

3 jeux de données:

- JDR (Réaliste)
- JDF (Factice)
- JDA (Authentique)

form	label	begin	end
+33 0134354919	Phone_Number	159	173
03.18.38.37.37	Phone_Number	124	138
+ 03 81 20 48 27	Phone_Number	137	153
01.75.88.25.30	Phone_Number	68	82
+ 33 01 77 83 74 05	Phone_Number	142	161
+33 0365962110	Phone_Number	68	82

Étiquettes simple

```
{'identifiant': 510,
 'text': 'Odette Camus\nData engineer - Libera\nXFactium\nAdresse : 25 bou
 levard de Prague 79000 Niort\nCedex 01 CS 45213\n+ 33 0529308213 / odettec
 amus@outlook.com\nLinkedin\nlibera.com',
 'annotations': [{'form': 'Odette', 'label': 'Human', 'begin': 0, 'end':
 6},
 {'form': 'Camus', 'label': 'Human', 'begin': 7, 'end': 12},
 {'form': 'Data', 'label': 'Function', 'begin': 13, 'end': 17},
 {'form': 'engineer', 'label': 'Function', 'begin': 18, 'end': 26},
```

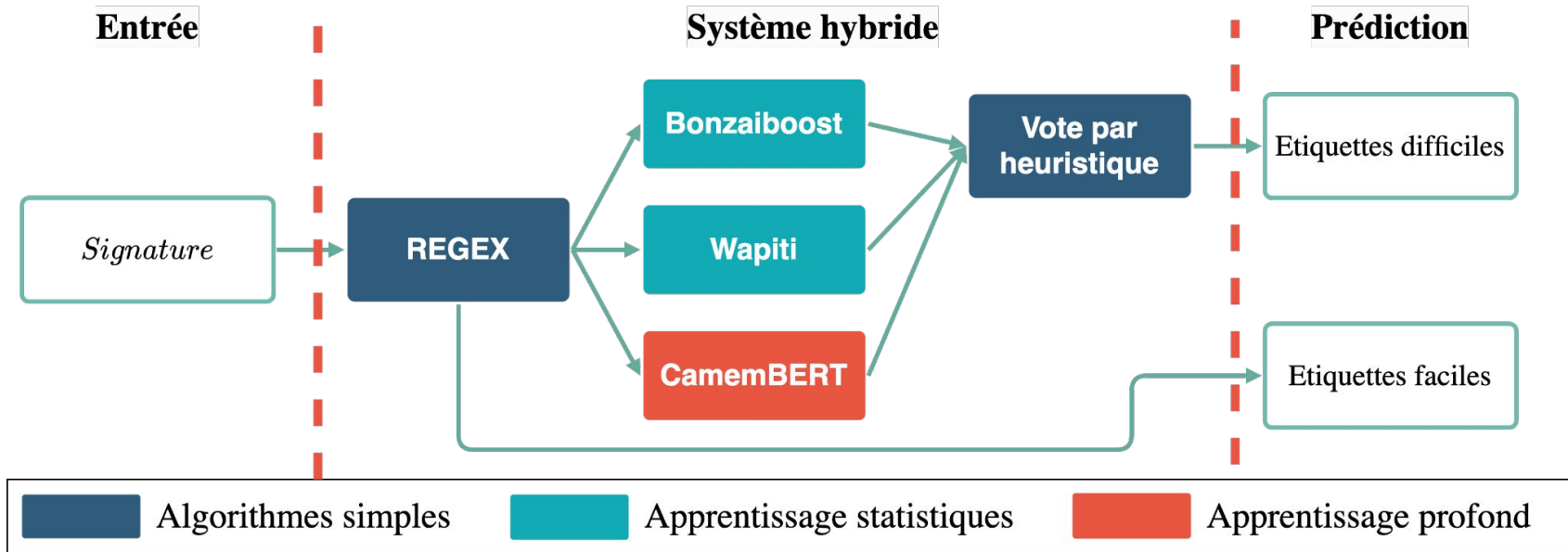
Figure 1 : Extraction d'un exemple dans JDR

```
{
  "form": "Parc",
  "begin": 31,
  "end": 35,
  "label": "Organization"
},
```

```
{
  "form": "des",
  "begin": 36,
  "end": 39,
  "label": "Organization"
},
```

```
{
  "form": "Guillaumes",
  "begin": 40,
  "end": 50,
  "label": "Organization"
},
```

Étiquettes difficiles



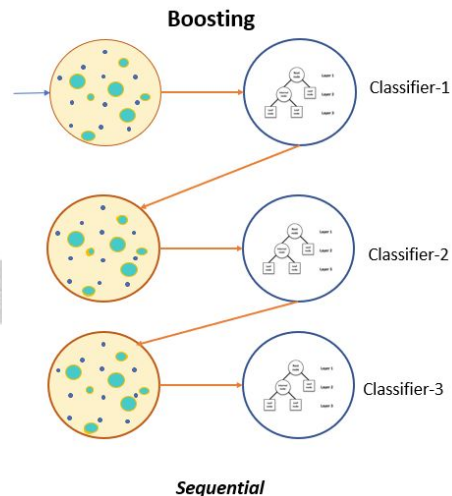
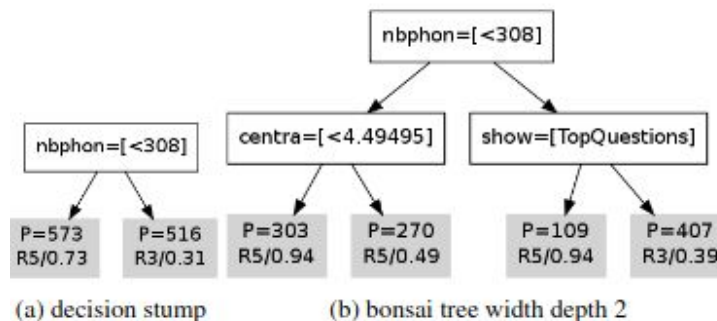
Bonzaiboost [1]

Bonsai = arbre peu profond

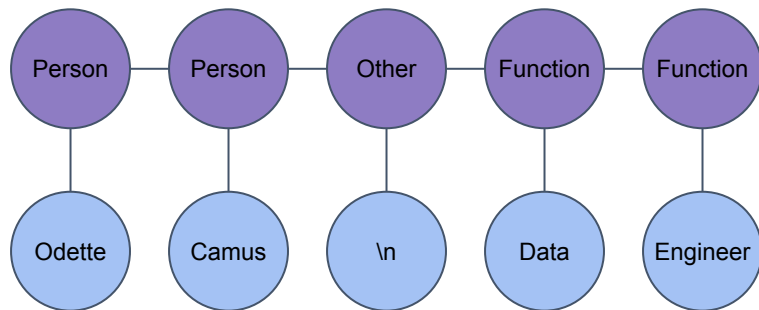
REGEX

`[\w\-\'\.]+\@([\w\-\.\.]+\.\w\w+)`

Chercheur à l'Inria,
thomas.simon2@inria.fr,
numéro de tél. ...

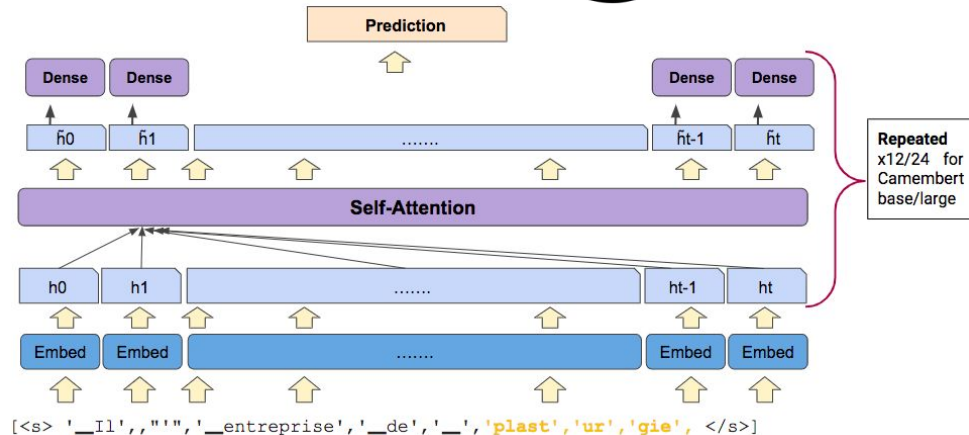


Wapiti [1]



Modèle probabiliste discriminatif, “champs conditionnels aléatoires”, avec **étiquettes** suivant l’hypothèse markovienne

CamemBERT [2]



Transformers, spécialisé sur corpus français

[1] Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical Very Large Scale CRFs.

[2] Martin, L., B. Muller, P. J. Ortiz Suárez, Y. Dupont, L. Romary, É. de la Clergerie, D. Seddah, et B. Sagot (2020).

CamemBERT : a tasty French language model.

Heuristique pour agréger les prédictions

“Est-ce que je détecte une entité?”

signatures

Vote

REGEX

oui

non

Vote
Majoritaire

Prédiction

Bonzaiboost

Wapiti

CamemBERT

Priorisation

CamemBERT

oui

non

Wapiti

oui

non

Bonzaiboost

Prédiction

Composants	F-score	Précision	Rappel
RegEx + Wapiti + Bonzaiboost	0.757	0.785	0.680
RegEx + Wapiti + Bonzaiboost + CamemBERT + vote	0.807	0.835	0.714
RegEx + Wapiti + Bonzaiboost + CamemBERT + vote avec heuristique de priorisation	0.832	0.800	0.778

- Le contexte est important pour généraliser sur JDA (BERT ajoute dans performance).
- La stratégie d'agrégation pertinente prend le risque d'accepter à tort certaines étiquettes mais permet de gagner plus de score.

Wapiti



Bonzaiboost



REGEX

CamemBERT



www.irisa.fr  @irisa_lab



Institut de Recherche en Informatique et Systèmes Aléatoires