

ETAPE

Système de détection d'Entités Nommées de l'IRISA

Christian Raymond

UEB - INSA de Rennes - IRISA - UMR 6074

18 Juin 2012

Approche générale

Approche supervisé

- robustesse sur transcription automatique
- Utilisation de CRF

Problématiques pour l'apprentissage

```
le  
<func.ind>  
|      <kind> maire </kind>  
|      de  
|      <loc.adm.town> Paris </loc.adm.town>  
</func.ind>
```

Approche générale

Approche supervisé

- robustesse sur transcription automatique
- Utilisation de CRF

Hiérarchie des entités

- Comment représenter la hiérarchie ?
- Doit on exploiter les sous classes ?

postulat sous classes pas généralisables

solution n'exploiter que la version la plus typée

Approche générale

Approche supervisé

- robustesse sur transcription automatique
- Utilisation de CRF

Imbrication des entités

- doit on exploiter la structure ?
- *a priori* oui, mais comment ?
- en cascade ? mais comment ?
- mais vraiment utile ?

postulat l'info de la structure est caché, mais on va la retrouver


observation pas d'imbrication d'EN de même type

solution toute entité est indépendante des autres

Définition du problème de classification

- ① Les labels sont les entités les plus typées
- ② On oublie la structure :
 - un classifieur par entité
 - réduction à 68 problèmes binaires
- ③ On reconstruit la structure :
 - On connaît la position relative d'une entité vs. une autre
 - entité de plus haut niveau se termine :
 - quand le classifieur le dit si possible
 - sinon attend que les entités imbriquées terminent
 - très basique, peut introduire des incohérences (pas mesuré)

Descripteurs utilisés



LABEL :	null	null	null	loc.adm.town	null	null	null
CLASSE :	ADV	NPMS	<unk>	NPSIG	NPSIG	NCMS	CAR
MOT :	Ici	Jacques	doutisoro	lomé	africa	numéro	un
POSITION :	-3	-2	-1	0	+1	+2	+3

Figure: Exemple d'étiquetage en entités nommées à partir des descripteurs de premier et second niveaux


premier niveau

- Les mots de la transcriptions

second niveau

- **ms** : résultat d'un étiquetage morpho-syntaxique
- **ap** : classe de généralisation (pays, villes, gentilés, ...)
- **mi** : mot « important »

Descripteurs utilisés



LABEL :	null	null	null	Loc.adm.town	null	null	null
CLASSE :	lci	PRENOM	<unk>	VILLE	NPSIG	numéro	un
MOT :	lci	Jacques	doutisoro	lomé	africa	numéro	un
POSITION :	-3	-2	-1	0	+1	+2	+3

Figure: Exemple d'étiquetage en entités nommées à partir des descripteurs de premier et second niveaux

premier niveau

- Les mots de la transcriptions

second niveau

- **ms** : résultat d'un étiquetage morpho-syntaxique
- **ap** : classe de généralisation (pays, villes, gentilés, ...)
- **mi** : mot « important »

Différentiel transcription man/auto

- Descripteur additionnel
 - longueur du mot courant
 - attribut numérique discrétisé
- Différences man/auto
 - 1 majuscules
 - 2 nombres : chiffres \longrightarrow lettres
 - 3 ponctuation

Discrétisation des attributs numériques

Méthode

- Méthode adaptée : [Fayyad and Irani, 1993]
- Méthode supervisé
- Chaque attribut est considéré indépendamment
- On construit un arbre de décision binaire entropique
- On utilise un critère spécial de stop sur le gain

[http://www.irisa.fr/texmex/people/raymond/Tools/
discretize4crf-3.1.tar.gz](http://www.irisa.fr/texmex/people/raymond/Tools/discretize4crf-3.1.tar.gz)

Exemple sur proba mot correct vs. incorrect

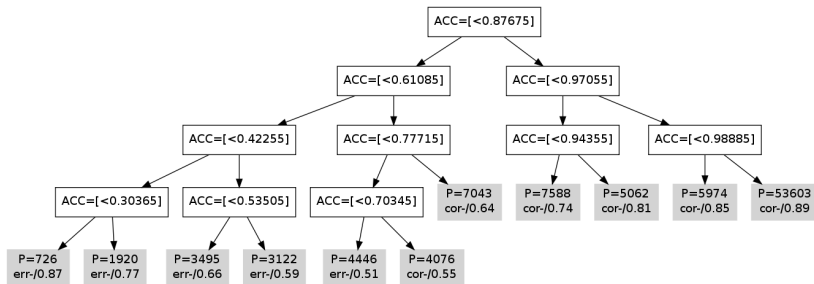


Figure: Exemple de discrétisation : feuilles=classes discrètes

Champs Conditionnels Aléatoires

$$p(\mathbf{e}|O) = \frac{1}{Z(O)} \exp\left(\sum_i \sum_k \lambda_k f_k(e_{i-1}, e_i, O, i)\right)$$

Où

$$f_k(e_{i-1}, e_i, O, i) = \begin{cases} 1 & \text{if } e_i = \text{loc.adm.town} \\ & \text{et } m_i = \text{à et } m_{i+1} = \text{Paris} \\ 0 & \text{otherwise} \end{cases}$$

et λ_k est le poids attribué à f_k lors de l'apprentissage

Résultats

	Man.	Rov.	s23	s24	s25	s30
eurecom	84.78	98.82	101.45	95.03	100.72	97.28
irisa	33.81	55.51	58.35	63.40	62.53	52.71
jouve	55.63	94.24	107.71	82.67	142.96	97.19
lif	43.58	69.54	74.55	71.93	85.60	69.24
limsi	36.44	67.16	68.57	67.73	75.02	60.44
lina-lium	62.76	76.45	80.84	77.97	82.71	76.63
synapse	42.89	68.65	74.93	70.77	86.10	66.23
tours	41.01	65.97	71.01	66.89	90.32	65.37



Fayyad, U. M. and Irani, K. B. (1993).

Multi-interval discretization of continuousvalued attributes for classification learning.

In *Thirteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 1022–1027. Morgan Kaufmann Publishers.