

Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement

Christian Raymond et Julien Fayolle

UEB - INSA de Rennes - IRISA - UMR 6074

20 Juillet 2010

Pourquoi ?

Documents multimédias numériques

- émission de radios, Journaux TV

↳ **sémantique forte contenue dans la parole**

Applications

- indexation, RI, structuration, résumé, traduction

↳ **Exploitation de la transcription automatique**

Reconnaissance d'Entités Nommées

- Flux brut de mots :
sémantique cachée

↳ Extraire du sens

Utilité pour la RI

jacques chirac

≠

jacques toubon bernadette
chirac

Comment ?

Texte propre ou transcriptions manuelles

- Modèles à base de grammaires formelles
- Définition manuelles des règles
- ↳ **Efficace** ($\approx 90\%$)
- ↳ **Très couteux à implémenter**
- ↳ **Inefficace sur sortie RAP** ($< 35\%$)

Transcriptions automatiques

- Modèles à base d'apprentissage automatique
- ↳ **Robuste sur sortie RAP** ($\approx 45\%$)
- ↳ **Nécessité d'un corpus d'exemples**

Robustesse pour Transcriptions automatiques

Plusieurs méthodes : avantages respectifs

Modèles d'apprentissage discriminant : *souplesse/performance*

- Champs condition. Aléatoires (CRF) : *adapté aux séquences*
- Machines à vecteur de support (SVM) : **large marge**

Modèle génératif : HMM

- Même paradigme que RAP
- Intégration avec le décodeur : traite des **graphes** de mots

Plusieurs méthodes : exploitation redondance

Différents systèmes \Rightarrow différentes vues

\leftrightarrow Exploitation à travers une opération de **fusion**

Définition du problème lors d'utilisation AA

entités		null		pers		loc		org	
étiquettes		O	pers-B	pers-I	loc-B	org-B	org-I	org-I	
mots		ici	jacques	doutisoro	lomé	africa	numéro	un	

Définition du problème lors d'utilisation AA

LABEL :	O	Pers-I	Pers-B	Loc-B	Org-B	Org-I	Org-I
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 »

Définition du problème lors d'utilisation AA

LABEL :	O	Pers-I	Pers-B	Loc-B	Org-B	Org-I	Org-I
CLASSE :	Ici	NPMS	<unk>	VILLE	NPSIG	numéro	un
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 »

Champs Conditionnels Aléatoires

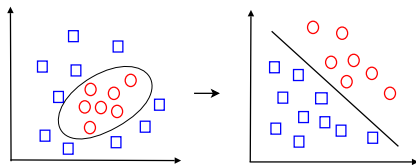
$$p(\mathbf{e}|O) = \frac{1}{Z(O)} \exp\left(\sum_i \sum_k \lambda_k f_k(e_{i-1}, e_i, O, i)\right)$$

Où

$$f_k(e_{i-1}, e_i, O, i) = \begin{cases} 1 & \text{if } e_i = \text{DATE} \\ & \text{et } m_i = \text{trente et } m_{i+1} = \text{octobre} \\ 0 & \text{otherwise} \end{cases}$$

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

Machines à vecteurs de support



- Problème non linéairement séparable
 - Projection dans un espace supérieur
- Trouve hyperplan qui maximise la marge
- Classifieur binaire → multi-classe
 - 1 vs. 1 ou 1 vs. tous
- Séquence → une classification par position

Modèle de Markov Caché

$$p(\hat{e}|A) \approx \operatorname{argmax}_{\mathbf{m}, \mathbf{c}, \mathbf{e}} p(A|\mathbf{m})p(\mathbf{m}|\mathbf{c}, \mathbf{e})p(\mathbf{c}, \mathbf{e})$$

Avec

A : observations acoustiques

e : séquence entités nommées

m : séquence classes (deuxième niveau de description)

m : séquence mots

Où

$p(A|\mathbf{m})$: donnée par système RAP

$p(\mathbf{m}|\mathbf{c}, \mathbf{e})$ probabilité d'émission du mot : estimée par Unigramme

$p(\mathbf{c}, \mathbf{e})$: probabilité jointe de transition : estimée par un Trigramme

Données expérimentales : Ester 2

Décomposition du corpus pour la tâche EN

Transcriptions d'émissions radiophoniques francophones
(France-Inter, France Info, RFI, RTM, France Culture, Radio Classique)

corpus	nombre d'heures	source
entraînement	60h	apprentissage Ester 1
	6h	développement Ester 2
test	6h	test Ester 2

Jeu d'Entités Nommées+Métriques d'évaluation

- 7 catégories principales : personne, fonction, organisation, lieu, production humaine, date et heure, montant
- Métriques : SER + F-mesure vs. Classifieur/Descripteur

Évaluation numéro 1

descripteurs	mot	mot+ms	mot+ms +ap	mot+ms +ap+mi
transcription	man	man	man	man
HMM	32.3 (0.77)	31.9 (0.77)	32.2 (0.77)	30.9 (0.78)
SVM	35.1 (0.77)	29.4 (0.80)	29.1 (0.81)	28.9 (0.81)
CRF	41.7 (0.72)	29.8 (0.79)	28.4 (0.80)	28.1 (0.80)

Évaluation numéro 1

descripteurs	mot	mot+ms	mot+ms +ap	mot+ms +ap+mi
transcription	man	man	man	man
HMM	32.3 (0.77)	31.9 (0.77)	32.2 (0.77)	30.9 (0.78)
SVM	35.1 (0.77)	29.4 (0.80)	29.1 (0.81)	28.9 (0.81)
CRF	41.7 (0.72)	29.8 (0.79)	28.4 (0.80)	28.1 (0.80)

Correction des données d'entraînement

Origine des erreurs

- Annotation des données d'entraînement non-homogène
- Train → Ester 1
- Dev → Ester 2
- Jeu d'EN + protocole légèrement différent

Procédure automatique de correction

- Train → mot+**ms**
- Dev → mot+**ms**+**ap**+**mi**
- Apprendre un CRF
- Train+Dev → mot+**ms**+**ap**+**mi**
- ré-annoter avec le CRF → **nouvelles annotations !**

Évaluation après correction

descripteurs	mot	mot+ms	mot+ms +ap	mot+ms +ap+mi
transcription	man	man	man	man

Évaluation **avant** correction

HMM	32.3 (0.77)	31.9 (0.77)	32.2 (0.77)	30.9 (0.78)
SVM	35.1 (0.77)	29.4 (0.80)	29.1 (0.81)	28.9 (0.81)
CRF	41.7 (0.72)	29.8 (0.79)	28.4 (0.80)	28.1 (0.80)

Évaluation **après** correction

HMM	27.3 (0.81)	29.6 (0.80)	29.1 (0.80)	26.6 (0.82)
SVM	32.4 (0.79)	27.4 (0.82)	26.9 (0.83)	26.6 (0.83)
CRF	36.2 (0.76)	24.8 (0.83)	23.4 (0.84)	22.8 (0.84)

Évaluation finale en SER selon la moulinette Ester 2

Comparaison des 3 systèmes avec les meilleurs systèmes Ester 2

ystème	HMM	SVM	CRF	best system manuel.	best system autom.
man	27.89	28.06	22.79	9.80	23.91
aut	59.44	59.83	53.49	66.22	56.79

Évaluation Oracle sur la fusion des systèmes : gain potentiel

(SVM+CRF)	(HMM+SVM+CRF)
50.40	45.80

Conclusion

- 3 systèmes robustes de reconnaissance EN
- gain potentiel sur combinaison
- méthode automatique de correction

Questions ?