TAL et Extraction d'Information

Isabelle Tellier

université Paris 3 - Sorbonne Nouvelle

Plan

1. Introduction : l'Extraction d'Information (EI)
2. Approches à base de règles
3. Approches par apprentissage automatique
4. Hybridations
5. Conclusion

Schéma général de la tâche d'EI

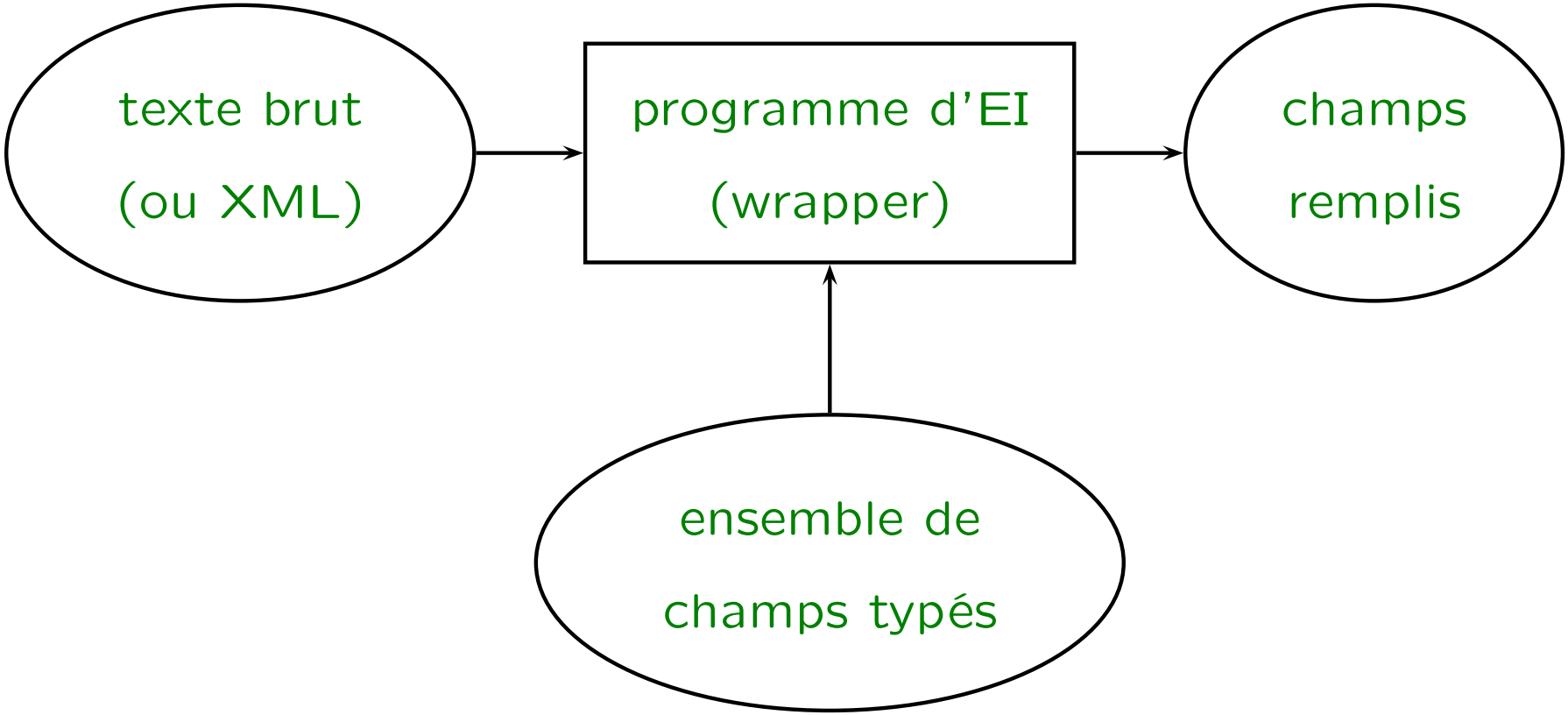


Figure 1: Schéma général de la tâche d'Extraction d'Information

1. L'Extraction d'Information

Origine et spécificités de la tâche

- origine : challenges MUC (1987-1998) organisés par la DARPA
- but : extraire des informations factuelles de données textuelles
- exemple (MUC 1992) : extraire de dépêches d'agences de presses sur des attentats des informations du type : date, lieu, nombre de victimes, auteur présumé/revendiqué...
- applications actuelles :
  - synthèse de textes de type journalistique (5W), de mails spécialisés (appels à soumissions de conférences...)
  - analyse de bibliographies (Google Scholar...), de CV...
  - analyse de descriptions de produits, petites annonces...

$\Longrightarrow$ l'EI vise à une distillation de l'information (McCallum)

Origine et spécificités de la tâche

- les données
  - leur nature textuelle est préservée : l'ordre des mots compte
  - XML/HTML : plusieurs lectures possibles (séquence de caractères, arbre, rendu visuel d'un navigateur)
- les champs typés à remplir
  - spécifiques pour chaque tâche
  - propriétés : obligatoire/facultatif, atomicité ("mot", "groupe de mots", (portion d'une ?) feuille d'un arbre...), multiplicité de la valeur (unaire/n-aire...)...
- la sortie attendue
  - format attendu variable (date, noms propres...)
  - unicité/multiplicité sur le texte

  $\Longrightarrow$ difficultés pour l'évaluation

Liens avec le TAL et la RI

- la plupart des champs sont de type entités nommées
  - noms propres : personne/lieu/organisation
  - quantités mesurables : date/valeur numérique
- l'évaluation se fait avec les même indicateurs que la RI : précision rappel, F-Mesure
- problèmes annexes, utilisations :
  - reconnaître les chaînes de co-références
  - entity linking : relier un nom et un identifiant dans une BD
  - identifier les relations prédicatives entre entités
  - peupler automatiquement une ontologie (par exemple pour les systèmes question/réponse), alimenter le Web sémantique

$\Longrightarrow$ très utile pour l'indexation, l'interrogation... de textes

Plan

1. Introduction : l'Extraction d'Information (EI)
2. Approches à base de règles
3. Approches par apprentissage automatique
4. Hybridations
5. Conclusion

2. Approches à base de règles

Propriétés

- règles de type expressions régulières écrites à la main, pour la reconnaissance des entités nommées en fonction de leur contexte
- gros usage de listes, dictionnaires...
- intérêt : lisibilité (jusqu'à un certain point)
- mais requiert une certaine expertise linguistique
- problème : grande évolutivité des noms, ambiguïtés...
- en général : bonne précision, mauvais rappel !
- exemple (démo) : Unitex

Plan

1. Introduction : l'Extraction d'Information (EI)
2. Approches à base de règles
3. Approches par apprentissage automatique
4. Hybridations
5. Conclusion

2. Approches à base de règles

Propriétés

- pour l'EI : apprentissage automatique supervisé
- via une reformulation du problème
- nécessite de disposer (de beaucoup) d'exemples annotés
  - pour l'apprentissage
  - pour l'évaluation
- en général : meilleur (surtout en rappel) que les règles

Plan

1. Introduction : l'Extraction d'Information (EI)
2. Approches à base de règles
3. Approches par apprentissage automatique
   - apprentissage symbolique
   - reformulation en une tâche de classification
   - un modèle d'annotation : les CRF
4. Hybridations
5. Conclusion

3. Approches par apprentissage automatique

Apprentissage symbolique : la PLI

- La PLI (Programmation Logique Inductive) vise à induire des règles logiques à partir d'exemples
  - exemples : $parent(Alain, Bart), parent(Bart, Carine)$
    $grand\_parent(Alain, Carine),$
  - règle : $\forall X, Y, Z$
    $grand\_parent(X, Z) \Longleftarrow parent(X, Y) \wedge parent(Y, Z)$
- règle pour EI : $\forall X$
  $nom\_pers(X) \Longleftarrow$
  $mot\_precedent("madame", X) \wedge commence\_maj(X)$
- le système Rapier (Califf & Mooney 03) fait de l'EI par PLI

3. Approches par apprentissage automatique

Apprentissage symbolique : inférence grammaticale

- L'IG vise à induire un modèle formel de langage (automate, grammaire, expression régulière...)
- à partir d'exemples (de séquences, d'arbres...) appartenant (ou non) à ce langage
- en théorie : de nombreux théorèmes
- en pratique : difficile à utiliser sur des données réelles
- en EI : des wrappers définis avec des grammaires d'arbres HTML appris par IG (Gilleron, Carme, Tommasi...)

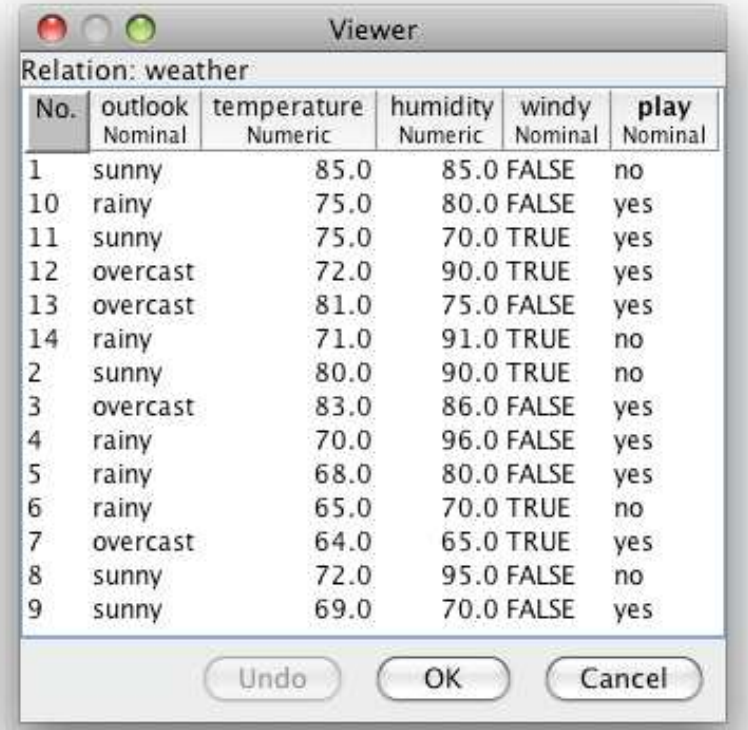
Plan

1. Introduction : l'Extraction d'Information (EI)
2. Approches à base de règles
3. Approches par apprentissage automatique
   - apprentissage symbolique
   - reformulation en une tâche de classification
   - un modèle d'annotation : les CRF
4. Conclusion

3. Approches par apprentissage automatique

Reformulation en une tâche de classification

- les données : un tableau de propriétés (attributs)
- le résultat recherché : une étiquette (valeur de la dernière colonne)
- il existe de très nombreuses techniques : NaiveBayes, arbres de décision, k-plus proches voisins, SVM... (cf. Weka)



3. Approches par apprentissage automatique

Reformulation en une tâche de classification

En	2016	*les*	*Jeux*	*Olympiques*	*auront*	*lieu*	*a*	*Rio*	*de*	*Janeiro*
	date		evt	evt				lieu	lieu	lieu

peut être codé en étiquetant :

- chaque séparateur indépendamment (risques d'ambiguités)

	En		2016		*les*		*Jeux*		*Olympiques*		...
O		D		F		D		M		F	

- chaque couple de séparateurs en 0/1 ou suivant un type (beaucoup d'appelés, peu d'élus)
- chaque token avec la norme BIO (Begin/In/Out)

En	2016	*les*	*Jeux*	*Olympiques*	*auront*	*lieu*	*a*	*Rio*	*de*	*Janeiro*
O	D-B	O	E-B	E-I	O	O	O	L-B	L-I	L-I

- chaque token avec la norme BILOU (BIO+Last/Unique)

En	2016	*les*	*Jeux*	*Olympiques*	*auront*	*lieu*	*a*	*Rio*	*de*	*Janeiro*
O	D-U	O	E-B	E-L	O	O	O	L-B	L-I	L-L

3. Approches par apprentissage automatique

Reformulation en une tâche de classification

unité	Maj ?	Chiffre ?	Ponct ?	étiquette
En	1	0	0	O
2016	0	1	0	D-B
,	0	0	1	O
les	0	0	0	O
Jeux	1	0	0	E-B
Olympiques	1	0	0	E-I
...				

Intérêt, limite

- efficace pour beaucoup de tâches
- limite : chaque étiquette (ligne) est indépendante des autres

3. Approches par apprentissage automatique

Reformulation en une tâche de classification

- une annotation = une suite de classifications
- on fixe un sens de parcours de la donnée
- on peut utiliser le résultat de classifications précédentes suivant ce sens de parcours

Exemple

En 2016 , les Jeux Olympiques ...

O

3. Approches par apprentissage automatique

Reformulation en une tâche de classification

- une annotation = une suite de classifications
- on fixe un sens de parcours de la donnée
- on peut utiliser le résultat de classifications précédentes suivant ce sens de parcours

Exemple

En	2016	,	les	Jeux	Olympiques	...
O	D-B					

3. Approches par apprentissage automatique

Reformulation en une tâche de classification

- une annotation = une suite de classifications
- on fixe un sens de parcours de la donnée
- on peut utiliser le résultat de classifications précédentes suivant ce sens de parcours

Exemple

En	2016	,	les	Jeux	Olympiques	...
O	D-B	O				

3. Approches par apprentissage automatique

Reformulation en une tâche de classification

- une annotation = une suite de classifications
- on fixe un sens de parcours de la donnée
- on peut utiliser le résultat de classifications précédentes suivant ce sens de parcours

Exemple

En	2016	,	les	Jeux	Olympiques	...
O	D-B	O	O			

3. Approches par apprentissage automatique

Reformulation en une tâche de classification

- une annotation = une suite de classifications
- on fixe un sens de parcours de la donnée
- on peut utiliser le résultat de classifications précédentes suivant ce sens de parcours

Exemple

En	2016	,	les	Jeux	Olympiques	...
O	D-B	O	O	E-B		

3. Approches par apprentissage automatique

Reformulation en une tâche de classification

- une annotation = une suite de classifications
- on fixe un sens de parcours de la donnée
- on peut utiliser le résultat de classifications précédentes suivant ce sens de parcours

Exemple

En	2016	,	les	Jeux	Olympiques	...
O	D-B	O	O	E-B	E-I	

3. Approches par apprentissage automatique

Reformulation en une tâche de classification

unité	Maj ?	Chiffre ?	Ponct ?	étiq. prec.	étiquette
En	1	0	0	-	O
2016	0	1	0	O	D-B
,	0	0	1	D-B	O
les	0	0	0	O	O
Jeux	1	0	0	O	E-B
Olympiques	1	0	0	E-B	E-I

Limite et solution

- champ de vision borné (markovien), fixé à l'avance
- si erreur au début du parcours : risque de propagation
- solution : essayer de trouver directement toutes les étiquettes !

Plan

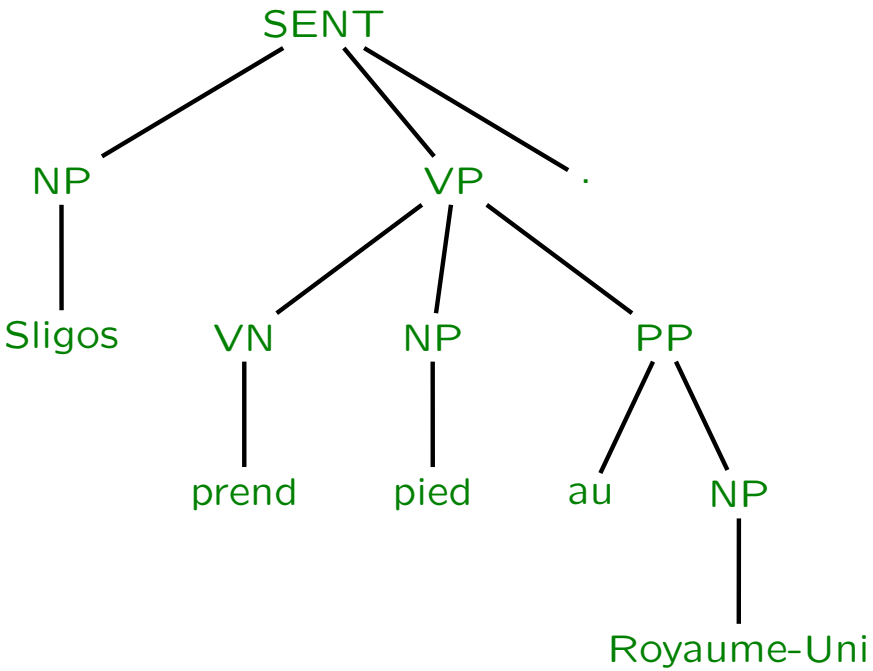
1. Introduction : l'Extraction d'Information (EI)
2. Approches à base de règles
3. Approches par apprentissage automatique
   - apprentissage symbolique
   - reformulation en une tâche de classification
   - un modèle d'annotation : les CRF
4. Hybridations
5. Conclusion

3. Approches par apprentissage automatique

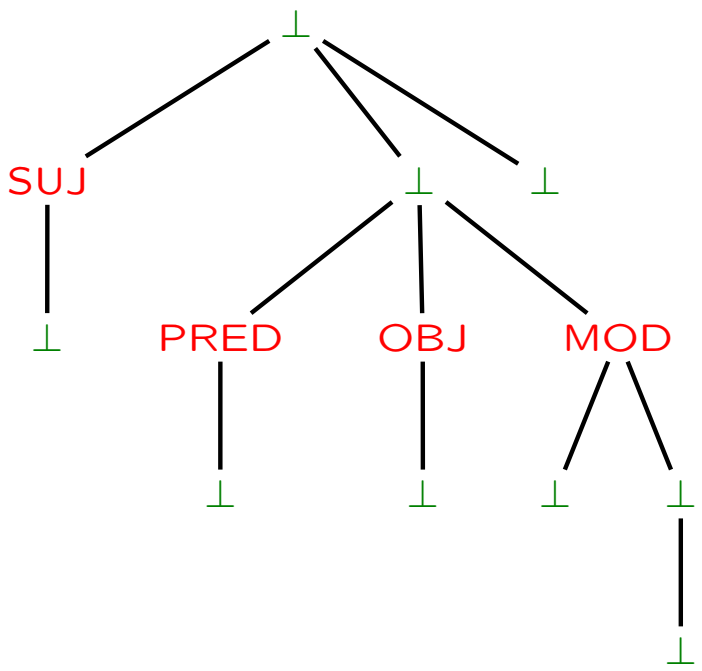
Notations de base

- x est une donnée complète, y est son annotation
- x et y ont la même structure
- ex. sur les séquences : $x = un\ chat\ dort$ et y=Det Nom Vintr
- ex. sur les arbres :

$x =$



y=



3. Approches par apprentissage automatique

Apprentissage automatique statistique

- on suppose qu'il existe une distribution de probabilité $p(y|x)$
- la forme du modèle p est fixée, à des paramètres près
- les deux problèmes qui se posent :
  - apprentissage : fixer les paramètres du modèle p à l'aide des couples (x, y)
  - annotation : une fois p fixé, pour tout nouveau x, trouver l'annotation y la plus probable, c'est-à-dire $\widehat{y} = argmax_y \; p(y|x)$

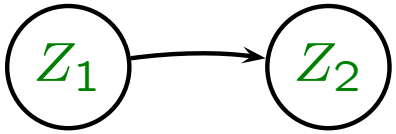
3. Approches par apprentissage automatique

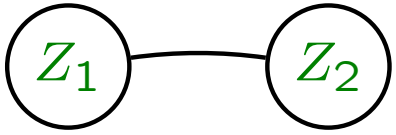
Apprentissage automatique statistique

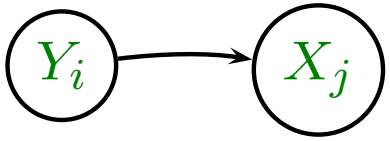
- une variable aléatoire est une variable pouvant prendre plusieurs valeurs données (cf. le dé...)
- on décompose x et y en des ensembles de variables aléatoires : $X = \{X_1, X_2, ..., X_n\}$ et $Y = \{Y_1, Y_2, ...Y_n\}$
- ex : les x sont des séquences de mots, les y leur étiquetage POS :
  - X_1 : variable dont les valeurs sont les 1ers mots des séquences x
  - Y_1 : variable dont les valeurs sont les 1ères étiquettes POS des séquences y, etc.
- intuition : il y a des dépendances entre les variables :
  - ex : si $X_i = le$, alors $Y_i =$Det ou $Y_i =$Pro
  - si en plus $Y_{i+1} =$Nom alors $Y_i =$Det...

3. Approches par apprentissage automatique

Modèles graphiques

- un modèle graphique définit les dépendances entre variables aléatoires par un graphe
- les variables aléatoires sont les nœuds du graphe
- 

: la valeur de Z_2 dépend de celle de Z_1
- 

: dépendances mutuelles entre Z_1 et Z_2
- dans un modèle graphique génératif, il y a des dépendances dirigées 

- exemples de modèle graphique génératif : les HMM, les PCFG
- Les CRF sont des modèles discriminants (= non génératifs)

3. Approches par apprentissage automatique

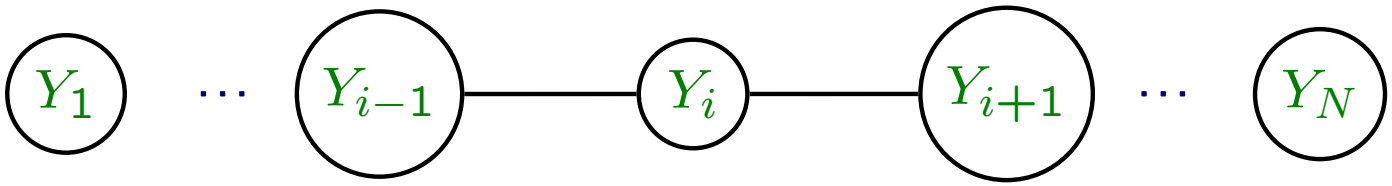
La formule des CRF

- proposition de (Lafferty, McCallum et Pereira 01) :

$$p(y|x) = \frac{1}{Z(x)} \prod_{c \in \mathcal{C}} \exp\left(\sum_k \lambda_k f_k(y_c, x, c) \right)$$

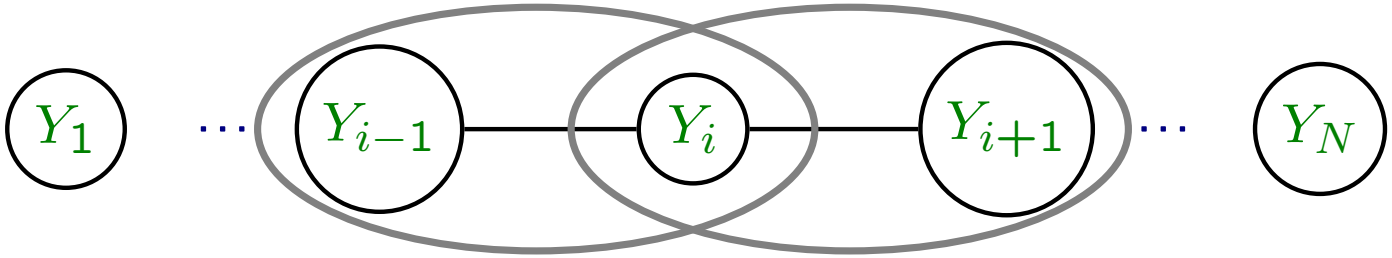
- $Z(x)$ un coefficient de normalisation
- $\mathcal{C}$ est l'ensemble des cliques du graphe sur Y
- chaque f_k est une feature donnée par l'utilisateur
- chaque λ_k est un poids initialement inconnu associé à f_k
- y_c : valeurs des variables de y sur la clique c

Les CRF "linéaires"



3. Approches par apprentissage automatique

Les CRF "linéaires"



- le graphe sur Y est une chaîne linéaire du 1er ordre
- implicite : chaque Y_i relié à chaque X_j
- les cliques sont les variables Y_i et les couples (Y_{i-1}, Y_i) (en gris)
- exemples de features $f_k(y_c, x, c)$ sur des séquences à la position i :
  * $f_k(y_{i-1}, y_i, x, i) = 1$ si $x_{i-1} \in \{la, une\}$ et $y_{i-1} = Det$ et $y_i = Nom$
    $= 0$ sinon
  * $f_{k'}(y_{i-1}, y_i, x, i) = 1$ si $\{M., Mme, Melle\} \cap \{x_{i-3}, ..., x_{i-1}\} \neq \emptyset$
    et $y_i = EN$
    $= 0$ sinon

3. Approches par apprentissage automatique

Les features : d'où viennent-elles ?

- plutôt que de définir des features, on définit des patrons
- un patron = une carte à trous (ou filtre coloré)
- une feature = une position du patron sur les données

unité	Maj ?	Chiffre ?	Ponct ?	étiquette
En	1	0	0	O
2016	0	1	0	D-B
,	0	0	1	O
les	0	0	0	O
Jeux	1	0	0	E-B
...				

$f(x, y_c)$ = (unité='En') ET (Maj=1) ET (Chiffre=0) ET (Ponct=0)
ET (y_c=O)

3. Approches par apprentissage automatique

Les features : d'où viennent-elles ?

- plutôt que de définir des features, on définit des patrons
- un patron = une carte à trous (ou filtre coloré)
- une feature = une position du patron sur les données

unité	Maj ?	Chiffre ?	Ponct ?	étiquette
En	1	0	0	O
2016	0	1	0	D-B
,	0	0	1	O
les	0	0	0	O
Jeux	1	0	0	E-B
...				

$f(x, y_c)$ = (unité='2016') ET (Maj=0) ET (Chiffre=1) ET (Ponct=0)
ET (y_c=D-B)

3. Approches par apprentissage automatique

Les features : d'où viennent-elles ?

- plutôt que de définir des features, on définit des patrons
- un patron = une carte à trous (ou filtre coloré)
- une feature = une position du patron sur les données

unité	Maj ?	Chiffre ?	Ponct ?	étiquette
En	1	0	0	O
2016	0	1	0	D-B
,	0	0	1	O
les	0	0	0	O
Jeux	1	0	0	E-B
...				

$f(x, y_c)$ = (unité=',') ET (Maj=0) ET (Chiffre=0) ET (Ponct=1)
ET (y_c=O)

3. Approches par apprentissage automatique

Les features : d'où viennent-elles ?

- plutôt que de définir des features, on définit des patrons
- un patron = une carte à trous (ou filtre coloré)
- une feature = une position du patron sur les données

unité	Maj ?	Chiffre ?	Ponct ?	étiquette
En	1	0	0	O
2016	0	1	0	D-B
,	0	0	1	O
les	0	0	0	O
Jeux	1	0	0	E-B

$f(x, y_c) =$ (unité='En') ET (Maj=1) ET (Chiffre=0) ET (Ponct=0)
(unité$_{+1}$='2016') ET (Maj$_{+1}$=0) ET (Chiffre$_{+1}$=1)
ET (Ponct$_{+1}$=0) ET (y_c=O) ET (y_{c+1}=D-B)

3. Approches par apprentissage automatique

Les features : d'où viennent-elles ?

- plutôt que de définir des features, on définit des patrons
- un patron = une carte à trous (ou filtre coloré)
- une feature = une position du patron sur les données

unité	Maj ?	Chiffre ?	Ponct ?	étiquette
En	1	0	0	O
2016	0	1	0	D-B
,	0	0	1	O
les	0	0	0	O
Jeux	1	0	0	E-B

$f(x, y_c) =$ (unité='2016') ET (Maj=0) ET (Chiffre=1) ET (Ponct=0)
(unité$_{+1}$=',') ET (Maj$_{+1}$=0) ET (Chiffre$_{+1}$=0)
ET (Ponct$_{+1}$=1) ET (y_c=D-B) ET (y_{c+1}=O)

3. Approches par apprentissage automatique

Les patrons, les features

- Les patrons peuvent prendre :
  - 1 ou 2 lignes successives sur la dernière colonne
  - n'importe quelle forme alilleurs
  - définis par une sorte d'expression régulière
- 1 patron = autant de features que de lignes où il s'applique
- les features permettent d'intégrer des connaissances externes
- 1 modèle CRF = 1 combinaison pondérée de features

3. Approches par apprentissage automatique

Ce qui est implémenté dans un CRF

- rappel de la formule, pour un graphe fixé

$$p(y|x) = \frac{1}{Z(x)} \prod_{c \in \mathcal{C}} \exp\left(\sum_k \lambda_k f_k(y_c, x, c) \right)$$

- problème de l'apprentissage :
  - les données : des (x, y) et les features fournies par l'utilisateur
  - problème : trouver les paramètres λ_k
  - méthode : chercher à maximiser la "log-vraisemblance" :

$$log(\prod_{(x,y) \in S} p(y|x)) = \sum_{(x,y) \in S} log\ p(y|x) (+penalisation)$$

    en cherchant où la dérivée (suivant tous les λ_k) s'annule...
- problème de l'annotation :
  - données : un CRF fixé, une nouvelle donnée x
  - problème : trouver le y qui maximise $p(y|x)$

3. Approches par apprentissage automatique

Les CRF linéaires

- nombreuses bibliothèques disponibles
- sont utilisables efficacement (millions de features pour Wapiti)
- ont été utilisés avec succès :
  - pour l'étiquetage POS (couplé avec segmentation)
  - pour le chunking, l'analyse syntaxique
  - pour la reconnaissance des entités nommées

Autre travaux avec un graphe non linéaire

- bibliothèque XCRF pour annoter des arbres XML
- problèmes de temps de calculs très longs

Plan

1. Introduction : l'Extraction d'Information (EI)
2. Approches à base de règles
3. Approches par apprentissage automatique
4. Hybridations
5. Conclusion

Intégrer des connaissances extérieures dans un CRF

- on peut les utiliser pour trier des étiquetages incohérents (bof...)
- il est facile de les intégrer dans les features
- on peut aussi en faire des exemples
- exemple : un lexique nous informe que :

  $en : prep, pro...$, $les : det, pro...$

unité	Liste-cat	...	étiquette
En	prep-pro		O
2016	-		D-B
,	ponct		O
les	det-pro		O
Jeux	nc		E-B
...			

Intégrer des connaissances extérieures

- on peut les utiliser pour trier des étiquetages incohérents (bof)
- il est facile de les intégrer dans les features
- on peut aussi en faire des exemples
- exemple : un lexique nous informe que :

  $en : prep, pro...$, $les : det, pro...$

unité	prep	nc	det	pro	...	étiquette
En	1	0	0	1		O
2016	-	-	-	-		D-B
,	0	0	0	0		O
les	0	0	1	1		O
Jeux	0	1	0	0		E-B
...						

Intégrer des connaissances extérieures

- on peut les utiliser pour trier des étiquetages incohérents (bof)
- il est facile de les intégrer dans les features
- on peut aussi en faire des exemples
- exemple : un lexique (ou un automate) nous informe que :
  4 chiffres = une date, les "Jeux Olympiques" sont un événement...

unité	date possible	evt possible	...	étiquette
En	0	0		O
2016	1	0		D-B
,	0	0		O
les	0	0		O
Jeux	0	1		E-B
Olympiques	0	1		E-I

Intégrer des connaissances extérieures

- on peut les utiliser pour trier des étiquetages incohérents (bof)
- il est facile de les intégrer dans les features
- on peut aussi en faire des exemples
- exemple : un lexique (ou un automate) nous informe que :
  4 chiffres = une date, les "Jeux Olympiques" sont un événement...

unité	date possible	evt possible	...	étiquette
En	0	0		O
2016	B	0		D-B
,	0	0		O
les	0	0		O
Jeux	0	B		E-B
Olympiques	0	I		E-I

Intégrer des connaissances extérieures

- on peut les utiliser pour trier des étiquetages incohérents (bof)
- il est facile de les intégrer dans les features
- on peut aussi en faire des exemples

  item exemple : un lexique (ou un automate) nous informe que :

  4 chiffres = une date, les “Jeux Olympiques” sont un événement...

unité	étiquette
Jeux	E-B
Olympiques	E-I

Plan

1. Introduction : l'Extraction d'Information (EI)
2. Approches à base de règles
3. Approches par apprentissage automatique
4. Hybridations
5. Conclusion

5. Conclusion

Intérêts des CRF

- outil très générique, neutre linguistiquement et efficace
- a fait ses preuves sur des tâches standards
- traitement homogène possible (via les features) d'informations de natures diverses : attributs, liens items/étiquettes et entre étiquettes
- exploitation possible de diverses manières de ressources externes
- combine connaissances symboliques locales (features) et lissage statistique global
- résultats (partiellement) interprétables

5. Conclusion

Utiliser l'apprentissage automatique ne dispense pas de réfléchir

- la tâche relève-t-elle de la classification/annotation ?
- dispose-t-on d'assez d'exemples annotés ?
- quel logiciel choisir ? quel protocole ?
- quels sont les attributs et patrons pertinents ?
- quelles sont les ressources exploitables, comment ?
- quels sont les paramètres du logiciel (pénalisation, descente de gradient...) pertinents ?
- quel critère de succès privilégier ?
  - efficacité en F-mesure
  - efficacité en temps de calcul
  - efficacité en nombres d'exemples requis
  - efficacité en robustesse sur d'autres données
  - lisibilité du résultat

5. Conclusion

L'Extraction d'Information

- tâche générale très utile
- fait le pont entre l'approche humaine d'un texte (compréhension) et son exploitation informatique (transformation en une BD)
- en interactions avec de nombreuses autres
  - TAL : annotation POS, reconnaissance/typage des EN, lien avec la syntaxe...
  - RI : indexation, interrogation, systèmes Q/R...
  - apprentissage automatique : apprentissage symbolique, classification, annotation...
- problèmes actuels :
  - entity linking, relations entre entités
  - entités nommées structurées