

Développement d'un système de reconnaissance d'entités nommées pour la campagne d'évaluation ETAPE

Christine Jacquin, Emmanuel Morin, **Mohamed Hatmi** & Sylvain Meignier

Atelier de clôture ETAPE

18 juin 2012

But :

Délimiter et catégoriser les entités nommées en respectant le guide d'annotation Quaero :

- 7 types principales et 32 sous-types
- pas de prise en compte des composants (e.g *name.first* et *name.last* pour le type *pers.ind*)

Approche :

Approche statistique basée sur le modèle de champs conditionnels aléatoires (CRF)

Méthodologie

Étapes :

- préparation des données d'apprentissage
- apprentissage du modèle
- développement des règles de post-traitement

Corpus d'apprentissage ETAPE

Traitement des balises imbriquées :

- `<func.ind> président du <org.ent> Front National </org.ent> </func.ind>` → `<func> président du </func> <org> Front National </org>`
- `<time.date.rel> depuis <amount> 21 ans </amount> </time.date.rel>` → `<time.date.rel> depuis </time.date.rel> <amount> 21 ans </amount>`
- `<org.adm> <loc.adm.nat> France </loc.adm.nat> </org.adm>` → `<org.adm> France </org.adm>`

Méthodologie : préparation des données d'apprentissage

Normalisation du corpus d'apprentissage :

- conversion des formes numériques dans leur forme textuelle
- suppression de la ponctuation
- conversion en minuscules
- conversion des formes particulières telles que "%" en "pourcent" et "." en "point" dans les URLs uniquement

Méthodologie : préparation des données d'apprentissage

Conversion en format BIO (Begin, Inside, outside) :

- transformation du corpus d'apprentissage sous forme des couples (x, y) :
 - x correspond au mot lui même ainsi que ses caractéristiques
 - POS-tags et présence de majuscule (0 ou 1) pour les transcriptions manuelles
 - POS-tags pour les transcriptions automatiques
 - y correspond à l'étiquette et à la position du mot dans l'entité nommée
- un mot par ligne
- phrases séparées par une ligne vide

Méthodologie : préparation des données d'apprentissage

Conversion en format BIO (Begin, Inside, outside) :

lundi NMS 0 time.date.abs-b
sept CHIF 0 time.date.abs-i
décembre NMS 0 time.date.abs-i

deux CHIF 0 amount-b
incendies NMP 0 amount-i
cette DETFS 0 time.hour.rel-b
nuit NFS 0 time.hour.rel-i
en PREP 0 0
région NFS 0 loc.adm.reg-b
parisienne AFS 0 loc.adm.reg-i

Méthodologie : préparation des données d'apprentissage

Étiquetage en parties du discours (POS) :

- utilisation du logiciel libre LIA_TAGG (Béchet, 2006)
- enrichissement du vocabulaire de base de LIA_TAGG avec :
 - 30 300 noms de personne
 - 18 700 noms d'organisation et d'entreprise
 - 62 600 noms de lieu

Méthodologie : entraînement des modèle

Entraînement des modèle :

- modèles entraînés avec l'implémentation des CRF
- utilisation du logiciel libre CRF++
<http://crfpp.sourceforge.net/>
 - fonctions booléennes générée en tenant compte des mots situés dans une fenêtre $[-2, +2]$ autour de la position courante

Méthodologie : post-traitement

Post-traitement :

- regroupement d'entités : règles manuelles

\$fonction \$article-min \$organisation → \$fonction
et le résultat de son application :

<func> président du </func> <org> Front National
</org> → <func.ind> président du
<org.ent> Front National </org.ent> </func.ind>

- double annotation de certaines entités : listes de noms propres

<org.adm> France </org.adm> → <org.adm>
<loc.adm.nat> France </loc.adm.nat> </org.adm>

Évaluation

Résultats : 38,0% de SER et 68% de F-mesure sur les transcriptions manuelles

- des résultats satisfaisants pour les catégories "classiques" :
 - pers.ind : F-mesure 86,8 (86,4/87,1)
 - loc.adm.town : F-mesure 73,4 (68,1/79,57)
 - loc.adm.nat : F-mesure 90,8 (88,4/93,5)
 - func.ind : F-mesure 60,9 (75,9/50,9)
 - org.adm : F-mesure 54,3 (71,8/43,7)
- des résultats moins bons pour les autres catégories :
 - loc.phys.geo : F-mesure 17,6 (75/10)
 - prod.object : F-mesure 6,2 (33,3/3,4)
 - prod.fin : F-mesure 52,6 (11,9/19,4)

Analyse des résultats

- problème d'annotation des expressions ne contenant pas des noms propres
 - prod.fin (hyp_count=19, ref_count=84, correct=10)
"Suites impériales", "Des éclairs" ...
 - prod.art (hyp_count=34, ref_count=87, correct=10)
"fond de garantie des calamités agricoles" ...
- difficulté de distinction entre certains sous-types peu fréquent dans le corpus d'apprentissage
 - loc.phys.hydro (hyp_count=2, ref_count=5, correct=0)
 - loc.add.phys (hyp_count=0, ref_count=4, correct=0)
 - prod.other (hyp_count=0, ref_count=7, correct=0)

Merci de votre attention