

Détection automatique à partir du web des caractéristiques ontologiques et du type d'une Entité Nommée

Ludovic Bonnefoy
ludovic.bonnefoy@ismart.fr

iSmart - Université d'Avignon (LIA)
Michel Benoit - Patrice Bellot

Mercredi 30 juin 2010

Problématique

La reconnaissance d'Entités Nommées

Définition

Etat de l'art

Entity Ranking à TREC

TREC

Entity Ranking

Méthode

Questions-réponses

Catégorisation des ENs candidates

Résultats

Perspectives

Problématique

- ▶ Déterminer si deux *entités nommées* sont de même type (représentent le même concept)
ex : *Nikon d3000* et *canon 550d*
- ▶ Etude de leur environnement
- ▶ Indépendant de la langue

Entité Nommée (EN)

Définition courante

- ▶ Unité d'information (ou textuelle)
- ▶ Représente une personne, une organisation, une date, etc.

Intérêt

- ▶ Filtre et amélioration des performances
- ▶ Systèmes de questions-réponses
- ▶ Ex : Quelles sont les villes à visiter en France ?

Etat de l'art

Méthodes déductives

- ▶ Utilisation de patrons d'extraction
- ▶ "Mr. mot_première_lettre_capitale ..." \Rightarrow mot est une personne.

Méthodes inductives : semi-supervisées

- ▶ Bootstrapping
- ▶ Synonymes, mots de la même classe [Pasca 06]
- ▶ Relations syntaxiques [Cucchiarelli 01]
- ▶ etc.

Etat de l'art (2)

Méthodes inductives : supervisées

- ▶ Arbres de décision [Sekine 98]
- ▶ CRF [Mc Mallum & Li 03]
- ▶ SVM [Asahara & Matsumoto 03]
- ▶ etc.

Ressources

- ▶ Corpus (Penn TreeBank)
- ▶ Listes d'ENs (gazeeters)
- ▶ Taxonomies (DBPedia)
- ▶ etc.

Limites

- ▶ Ex : Quels sont les acheteurs potentiels d'Eurocopter ?
- ▶ Faible nombre de types d'ENs reconnus (5-200)
- ▶ Impossibilité de construire des ressources aussi précises que nécessaire
- ▶ Monde en perpétuelle évolution

TREC

- ▶ Campagne d'évaluation internationale en RI
- ▶ Créée en 1992
- ▶ Soutenue par le gouvernement américain (NIST)

Entity Ranking

Related Entity Finding

- ▶ Retrouver des ENs d'un type donné
- ▶ Satisfaisant une relation en texte libre avec une EN source
- ▶ Exemple de Topic :
EN : Michael Schumacher
Type : personne
Texte : Michael's teammates while he was racing in Formula 1

Corpus et Evaluation

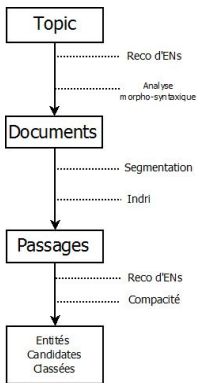
Corpus

- ▶ ClueWeb09
- ▶ 1 milliard de pages web
- ▶ 15To de données
- ▶ 4 langues (En, Fr, Es, Ch)

Evaluation

- ▶ Plusieurs mesures : P@R, MAP@R, nDCG@R
- ▶ 100 résultats maximum
- ▶ Cinquantaine de topics

Questions-réponses



Reconnaissance d'entités nommées

- ▶ Pers, Org et Loc : Stanford-NER
- ▶ Produits : Règles

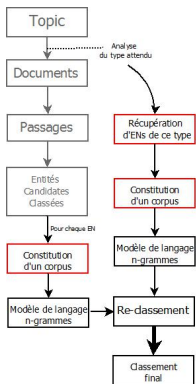
Analyse morpho-syntaxique

- ▶ TreeTagger

Compacité

$$\text{Comp}(EC_i) = \frac{1}{|QW|} \sum_{w \in QW} \frac{Z_j}{|R_j|+1}$$

Catégorisation des ENs candidates



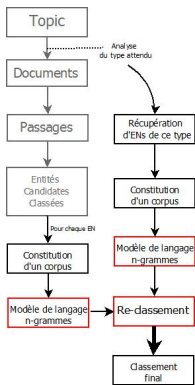
Collecte d'ENs de référence

- ▶ Du type attendu
- ▶ Utilisation de patrons d'extraction
Ex : "teammates including NP"
- ▶ Pasca 2004

Constitution de corpus

- ▶ Interrogation de Yahoo !
- ▶ Requête : EN
- ▶ Récupération des x premiers documents
- ▶ Corpus de référence :
constitué des documents récupérés pour chaque EN de référence.

Catégorisation des EN candidates (2)



Modèle de langage

- Calcul de n-grammes
- Lissage des distributions : Dirichlet

$$p(w|d) = \begin{cases} p_s(w|d) & \text{si } w \text{ est vu} \\ \alpha_d p(w|C) & \text{sinon} \end{cases}$$

$$p_s(w|d) = \frac{tf(c,d) + \mu p(w|C)}{\sum_{w' \in V} tf(w',d) + \mu}$$

$$\alpha_d = \frac{\mu}{\sum_{w \in V} tf(w,d) + \mu}$$

Divergence de Kullback-Leibler

- Distance entre deux distributions de probabilités
- Le modèle de référence et celui d'une EN candidate

$$DKL(P|Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)}$$

Résultats de la chaîne principale

Métrique	Méthode	Max	Median
P@10	.070	.235	.005

TAB.: Résultats, Référentiel TREC Entity Ranking 2009, Anglais. 13 équipes, 4 runs par équipe

Métrique	PERS	ORG	PROD	ALL
P@10	0,067	0,082	0,033	0,0700

TAB.: Catégories, Référentiel TREC Entity Ranking 2009, Anglais

Actuellement la catégorisation améliore les résultats d'environ 15%

Détection des caractéristiques

Première approche

- ▶ Analyse des différences de comportement d'unigrammes entre le modèle lié au corpus d'ENs d'un même type et un modèle du monde

Exemple

- ▶ Différence entre les probabilités des deux modèles d'unigrammes
- ▶ Type d'ENs : *camera*
- ▶ 20 premiers unigrammes :
measure, 200mm, \$599.95, 2000d, advertising, selector, symantec, focus, high-quality, digital, softcover, globally, select, seldom, differ, picture, softwaresecurity, decent-sized, high-contrast, ex-z300.

Merci de votre attention

Des questions ?

Métriques

Precision

$$\text{Precision} = \frac{\|\text{documents-pertinents-retrouves}\|}{\|\text{documents}_r\text{-trouves}\|}$$

Mean Average Precision (MAP)

$$\text{MAP@R} = \frac{\sum_{r=1}^R (P@r \cdot \text{rel}(r))}{\text{NbrDeDocumentARetrouver}}$$

normalized Discounted Cumulative Gain (nDCG)

$$\text{DCG@R} = \text{rel}_1 + \sum_{i=2}^R \frac{\text{rel}_i}{\log_2 i}$$

$$\text{nDCG@R} = \frac{\text{DCG@R}}{\text{IDCG@R}}$$