Extraction d'information

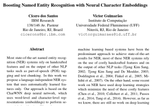
Anne-Laure Ligozat

màj: 2017

TAL & Web Sémantique

Documents textuels et bases de connaissances

article



Boosting Named Entity Recognition with Neural Character Embeddings

Cícero dos Santos
IBM Research
138/146 Av. Pasteur
Rio de Janeiro, RJ, Brazil
cicerons@br.ibm.com

Victor Guimarães
Instituto de Computação
Universidade Federal Fluminense (UFF)
Niterói, RJ, Rio de Janeiro
victorguimaraes@id.uff.br

Abstract

Most state-of-the-art named entity recognition (NER) systems rely on handcrafted features and on the output of other NLP tasks such as part-of-speech (POS) tagging and text chunking. In this work we propose a language-independent NER system that uses automatically learned features only. Our approach is based on the CharWNN deep neural network, which uses word-level and character-level representations (embeddings) to perform se-

machine learning based systems have been the predominant approach to achieve state-of-the-art results for NER, most of these NER systems rely on the use of costly handcrafted features and on the output of other NLP tasks (Tjong Kim Sang, 2002; Tjong Kim Sang and De Meulder, 2003; Doddington et al., 2004; Finkel et al., 2005; Milidiú et al., 2007). On the other hand, some recent work on NER have used deep learning strategies which minimize the need of these costly features (Chen et al., 2010; Collobert et al., 2011; Passos et al., 2014; Tang et al., 2014). However, as far as we know, there are still no work on deep learning

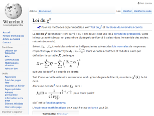
tweets

LIMSI NLP Group a retweeté
Marianna Apidianaki @MApidianaki · 24 oct.
Anders Søgaard will talk about "The language that wasn't on the map..." on Nov 22, 11:00-12:30 at @LimsiLab Don't miss it!
2 1

LIMSI NLP Group @LIMSI_NLP · 21 oct.
#NLProc for global mapping of health research dx.doi.org/10.1186/s12859... #ehealth
1

LIMSI NLP Group @LIMSI_NLP · 5 oct.
#job hiring #NLProc #textmining postdoctoral researcher- Greater Paris area @LimsiLab perso.limsi.fr/neveol/PostDoc

wikipédia



Wikipédia L'encyclopédie libre

Accueil
Portails thématiques
Article au hasard
Contact

Contribuer
Débuter sur Wikipédia
Aide
Communauté
Modifications récentes
Faire un don

Outils
Pages liées
Suivi des pages liées
Importer un fichier
Pages spéciales
Adresse permanente
Information sur la page
Élément Wikidata
Citer cette page

Imprimer / exporter

Article Discussion Lire Modifier Modifier le code

Loi du χ^2

Pour les méthodes expérimentales, voir Test du χ^2 et méthode des moindres carrés.

La **loi du χ^2** (prononcer « khi carré » ou « khi-deux ») est une loi à densité de probabilité. Cette loi est caractérisée par un paramètre dit degrés de liberté à valeur dans l'ensemble des entiers naturels (non nuls).

Soient $X_1, \ldots, X_k$, k variables aléatoires indépendantes suivant des lois normales de moyennes respectives μ_i et d'écart-type σ_i ; $Y_i = \frac{X_i - \mu_i}{\sigma_i}$ leurs variables centrées et réduites, alors par définition la variable X, telle que

$$X := \sum_{i=1}^{k} Y_i^2 = \sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i} \right)^2$$

suit une loi du χ^2 à k degrés de liberté.

Soit X une variable aléatoire suivant une loi du χ^2 à k degrés de liberté, on notera $\chi^2(k)$ la loi de X.

Alors une densité[1] de X notée f_X sera :

$$f_X(t) = \frac{1}{2^{\frac{k}{2}}\Gamma(\frac{k}{2})} t^{\frac{k}{2}-1} e^{-\frac{t}{2}}$$ pour tout t positif

où Γ est la fonction gamma.

L'espérance mathématique de X vaut k et sa variance vaut $2k$.

Sommaire [masquer]

page web

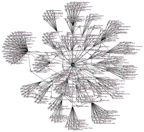
Six mois avec sursis pour des faits contestés avec méthode. Nicolas Jounin, jeune sociologue de 35 ans, a été condamné jeudi 3 novembre à six mois de prison avec sursis par le tribunal de Bobigny, qui l'a jugé coupable d'avoir frappé un policier au printemps dernier lors d'une manifestation contre la loi El Khomri. Nicolas Jounin a toujours démenti les faits, et notamment lors de l'audience le 6 octobre où son avocat avait pointé plusieurs contradictions dans les témoignages des policiers.

LIRE AUSSI
- Violences policières: ce que révèlent les cas Vadot et Jounin PAR CHRISTOPHE GUEUGNEAU
- Se disant victime de violences policières, un enseignant de Paris-1 porte plainte PAR CHRISTOPHE GUEUGNEAU
- Un sociologue en comparution immédiate PAR JOSEPH CONFAVREUX

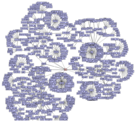
Le sociologue était accusé d'avoir bousculé et fait tomber un policier, puis de lui avoir asséné un coup de pied sur le casque, le 28 avril au matin. Ce jour-là, des syndicalistes s'étaient donné rendez-vous à Gennevilliers pour distribuer des tracts et inciter les travailleurs, du port notamment, à rejoindre le mouvement contre la loi El Khomri. Après un face-à-face avec des policiers, ils décident de prendre le métro pour se rendre Carrefour Pleyel à Saint-Denis, et rejoindre ainsi une assemblée générale interprofessionnelle prévue à 10 heures à la Bourse du travail. Une manifestation, déclarée, est prévue l'après-midi même au départ de la place Denfert-Rochereau à Paris.

Les faits reprochés à Nicolas Jounin se seraient déroulés Carrefour Pleyel. Selon le récit policier, un groupe aurait chargé un cordon positionné pour les empêcher de rejoindre la Bourse du travail. C'est à cette occasion que Nicolas Jounin aurait foncé sur un policier

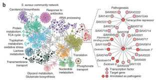
…s depuis www.mediapart.fr...



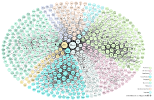
wikidata



wordnet

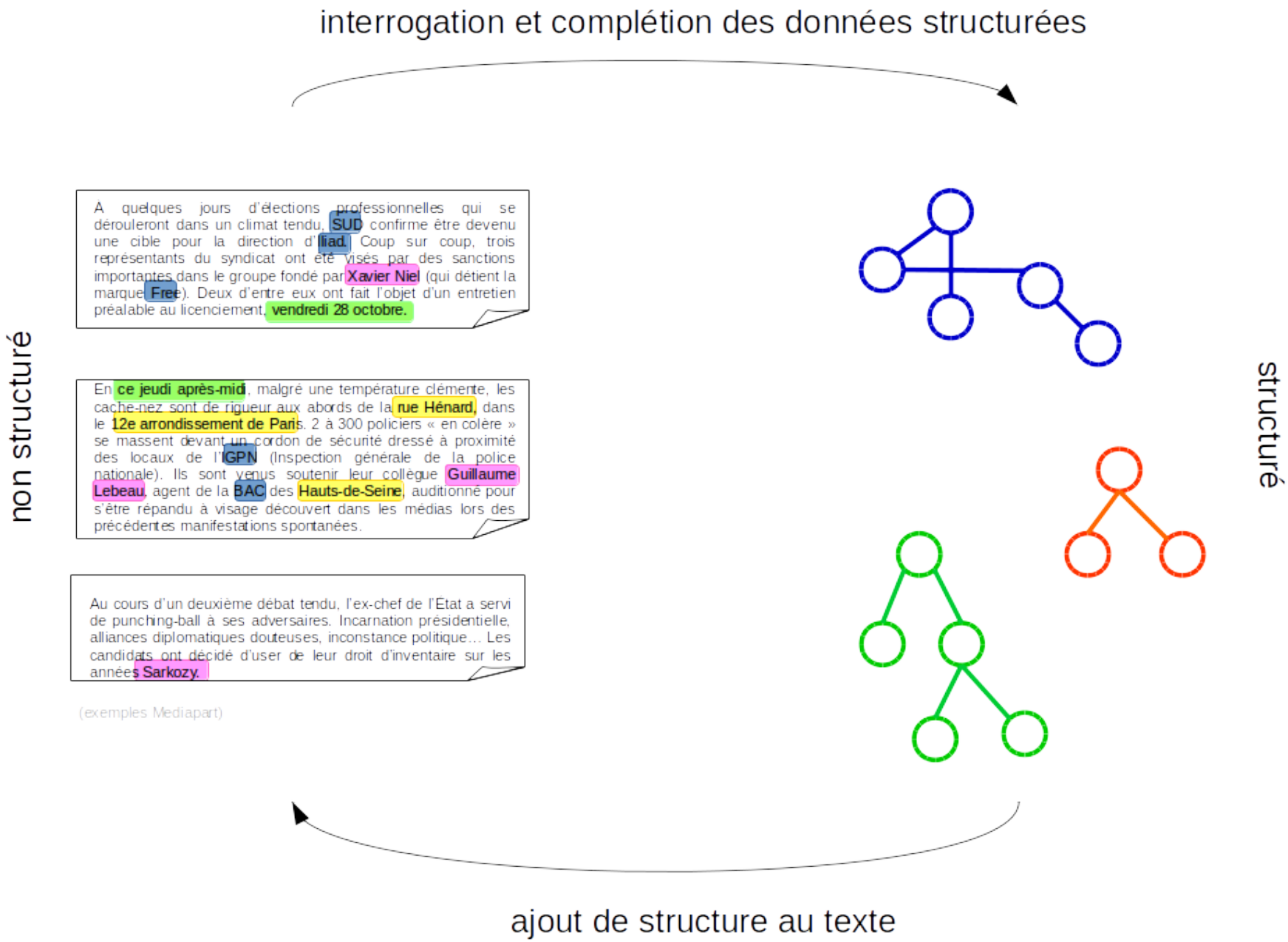


gene ontology



linking open data

Documents textuels et connaissances structurées



Extraction d'information/de connaissances

Objectifs

- compréhension ciblée de textes

→ produire une représentation structurée de l'information pertinente

- base de données relationnelle
- base de connaissances

→ raisonnement et inférence

Objectif

Résumé compréhensible par un programme

Involvement of Tumor Necrosis Factor Receptor-associated Protein 1 (TRAP1) in Apoptosis Induced by β-Hydroxyisovalerylshikonin*

Résumé textuel :
résumé pour humains

Subject	Relation	Object
p53	**is_a**	protein
Bax	**is_a**	protein
p53	has_function	apoptosis
Bax	has_function	induction
apoptosis	involved_in	cell_death
Bax	is_in	mitochondrial outer membrane
Bax	is_in	cytoplasm
apoptosis	related_to	caspase activation
...	...	...

Extraction de connaissances
structurées : résumé pour machines

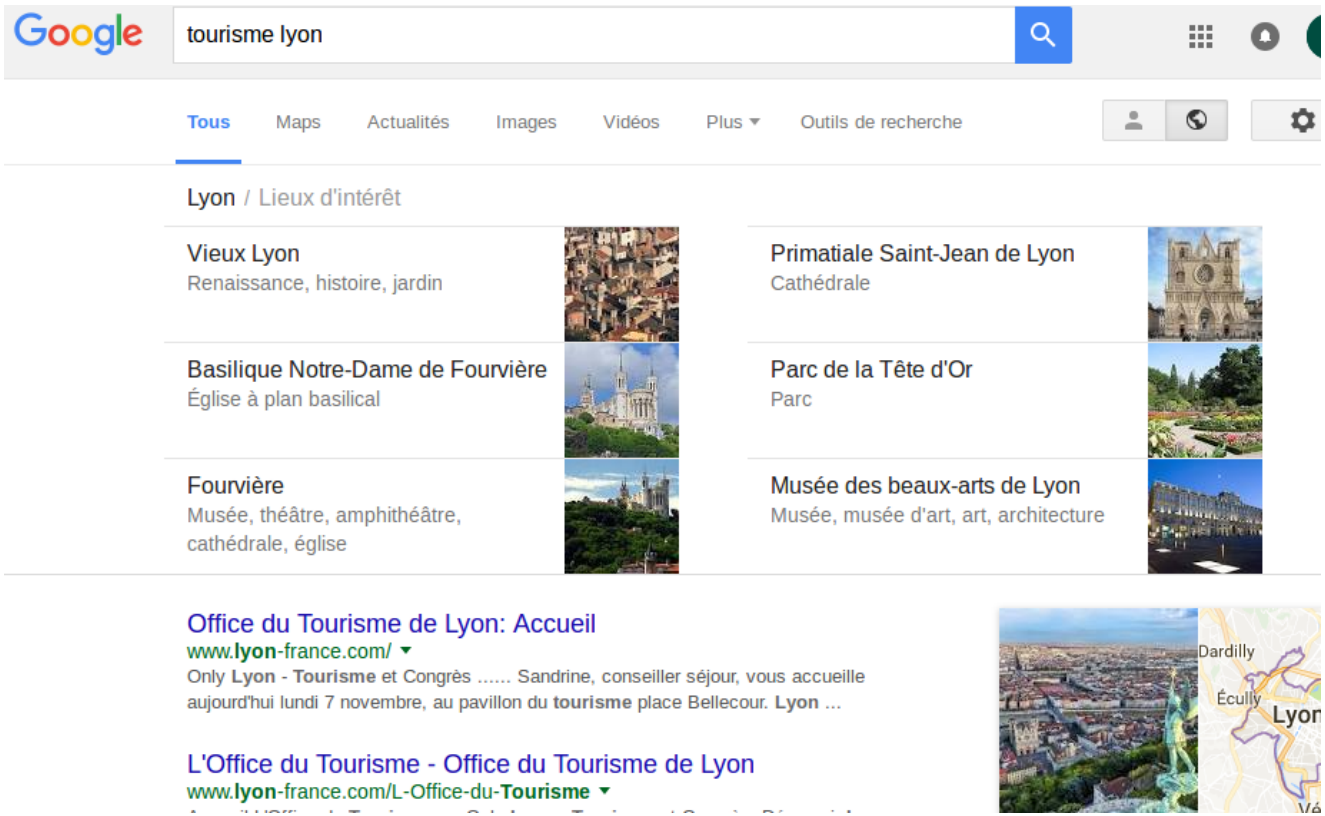
Compréhension "ciblée" ?

"The Lawrence Livermore National Laboratory (LLNL) in Livermore, California is a scientific research laboratory founded by the University of California in 1952."

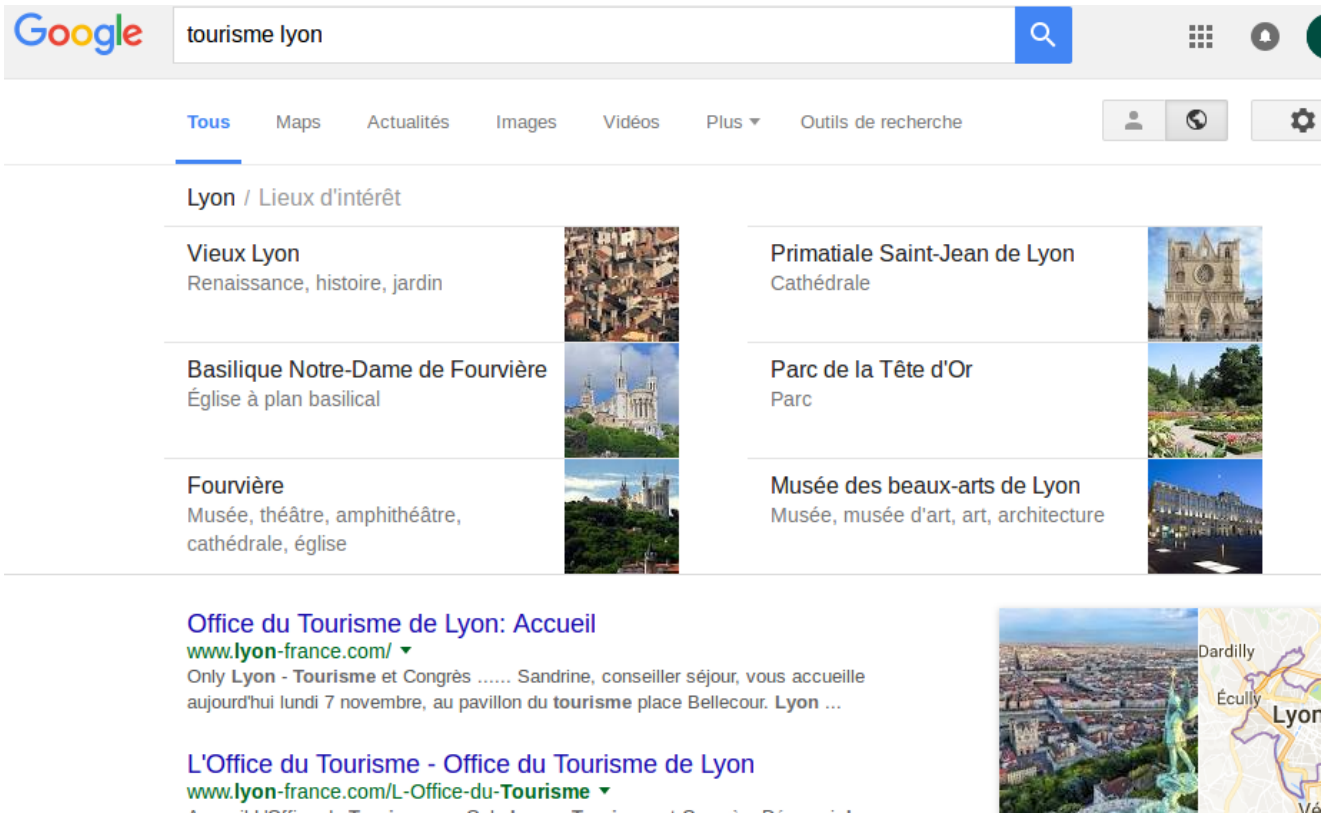
LLNL EQ Lawrence Livermore National Laboratory
LLNL LOC-IN *California*
Livermore LOC-IN *California*
LLNL IS-A scientific research laboratory
LLNL FOUNDED-BY University of California
LLNL FOUNDED-IN 1952

Applications de l'EI de la vie courante

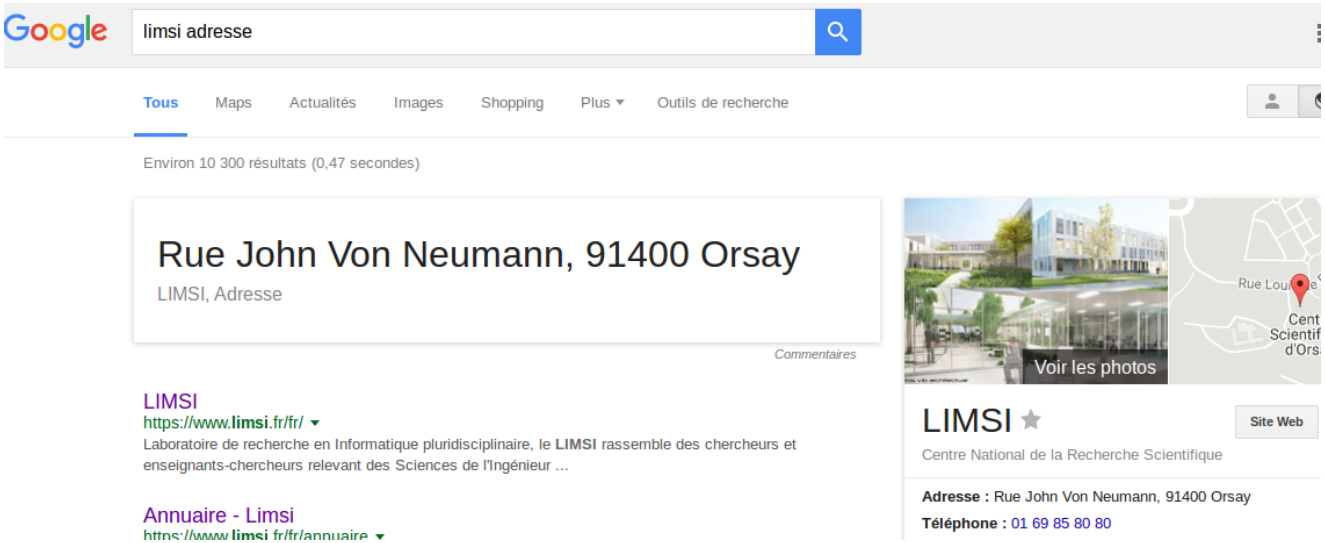
Applications de l'EI de la vie courante



Applications de l'EI de la vie courante



Applications de l'EI de la vie courante



Domaines d'application

- domaine général
- digital libraries (google scholar, citeseer)
- bioinformatique
- analyse des brevets...

Base de connaissances

Base de connaissances RDF

Ensemble de faits

- fait = sujet, prédicat, objet
  - ressources = entités, concrètes ou abstraites
  - propriétés = relations, comme height pour une Person

Exemple de triplet

<http ://dbpedia.org/resource/J._K._Rowling>
<http ://dbpedia.org/ontology/notableWork>
<http ://dbpedia.org/resource/Harry_Potter>

Quelques exemples de KB

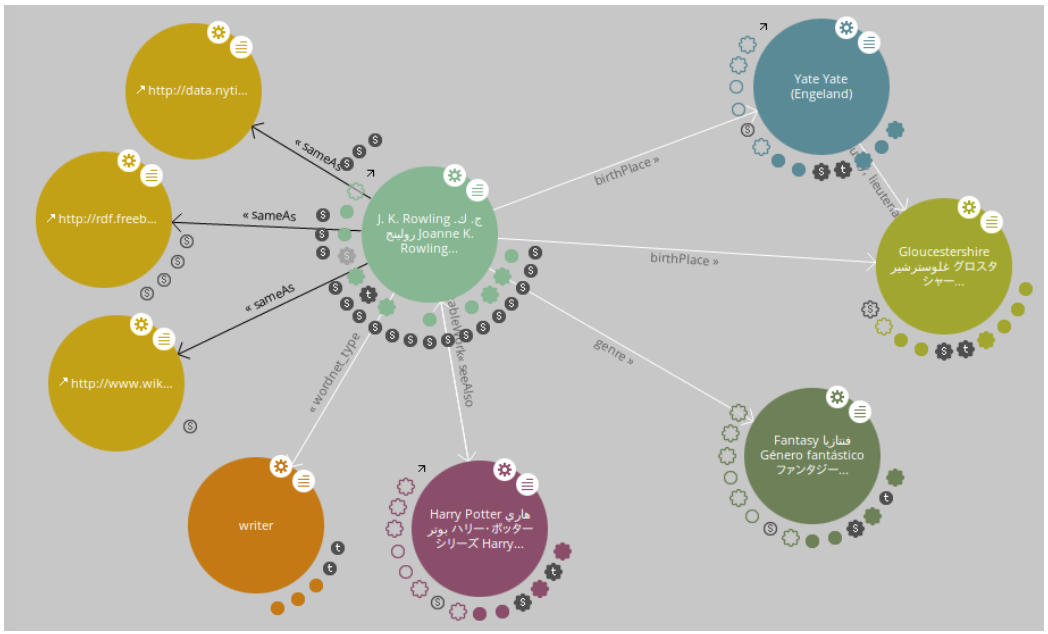
DBPedia

En résumé

- plus de 4 millions de ressources RDF (nov 2016)
- cœur = extraction des infobox
- ontologie partiellement intégrée (création manuelle)

Born	Joanne Rowling 31 July 1965 (age 51) Yate, Gloucestershire, England
Pen name	J. K. Rowling Robert Galbraith
Occupation	Novelist
Nationality	British
Education	Bachelor of Arts
Alma mater	University of Exeter
Period	1997–present
Genre	Fantasy, drama, young-adult fiction, tragicomedy, crime fiction
Notable works	*Harry Potter* series

→

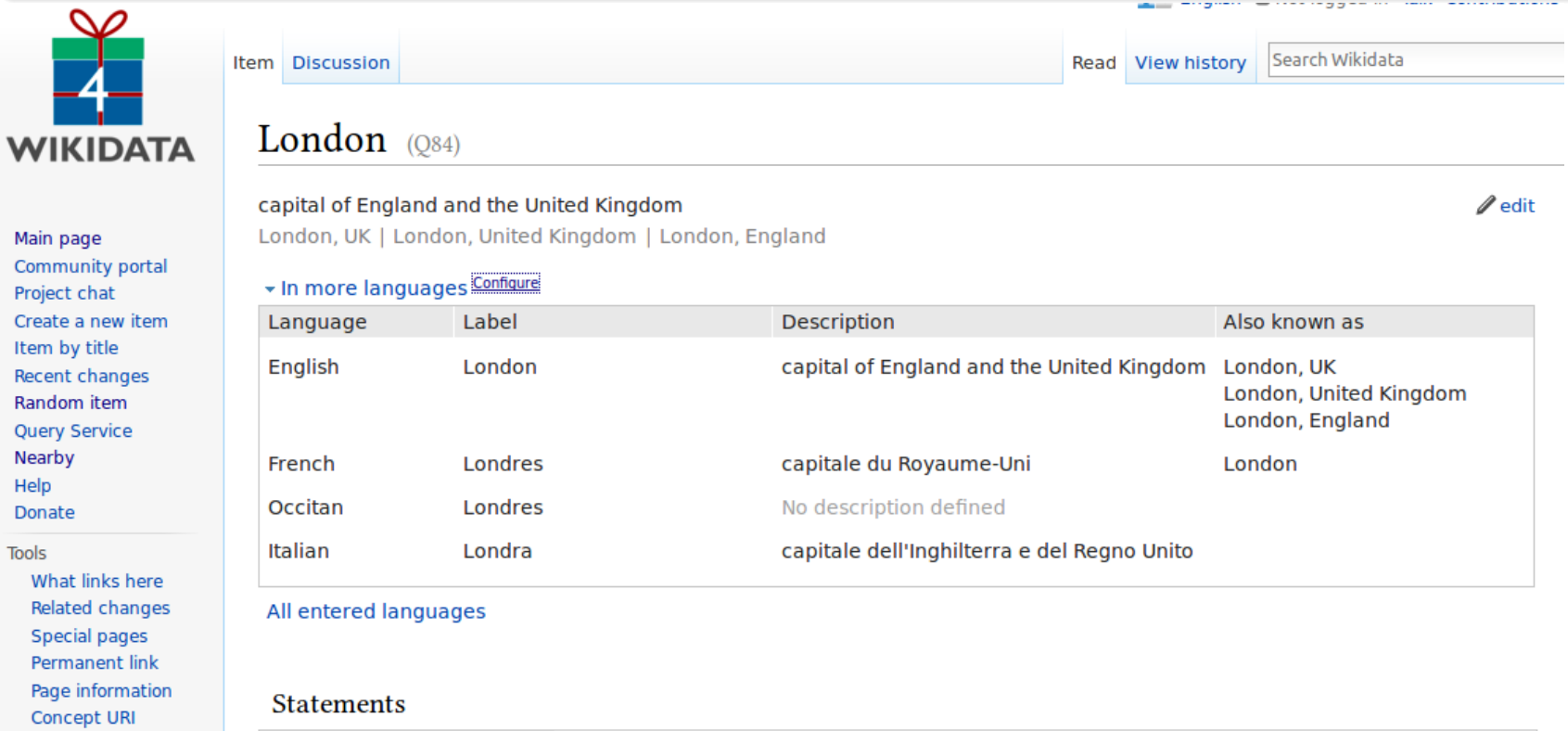


Quelques exemples de KB

Wikidata

En résumé

- tranfert partiel du contenu de Freebase
  - contenu collaboratif
  - import d'autres sources comme MusicBrainz
- plus de 20 millions d'items (nov 2016)

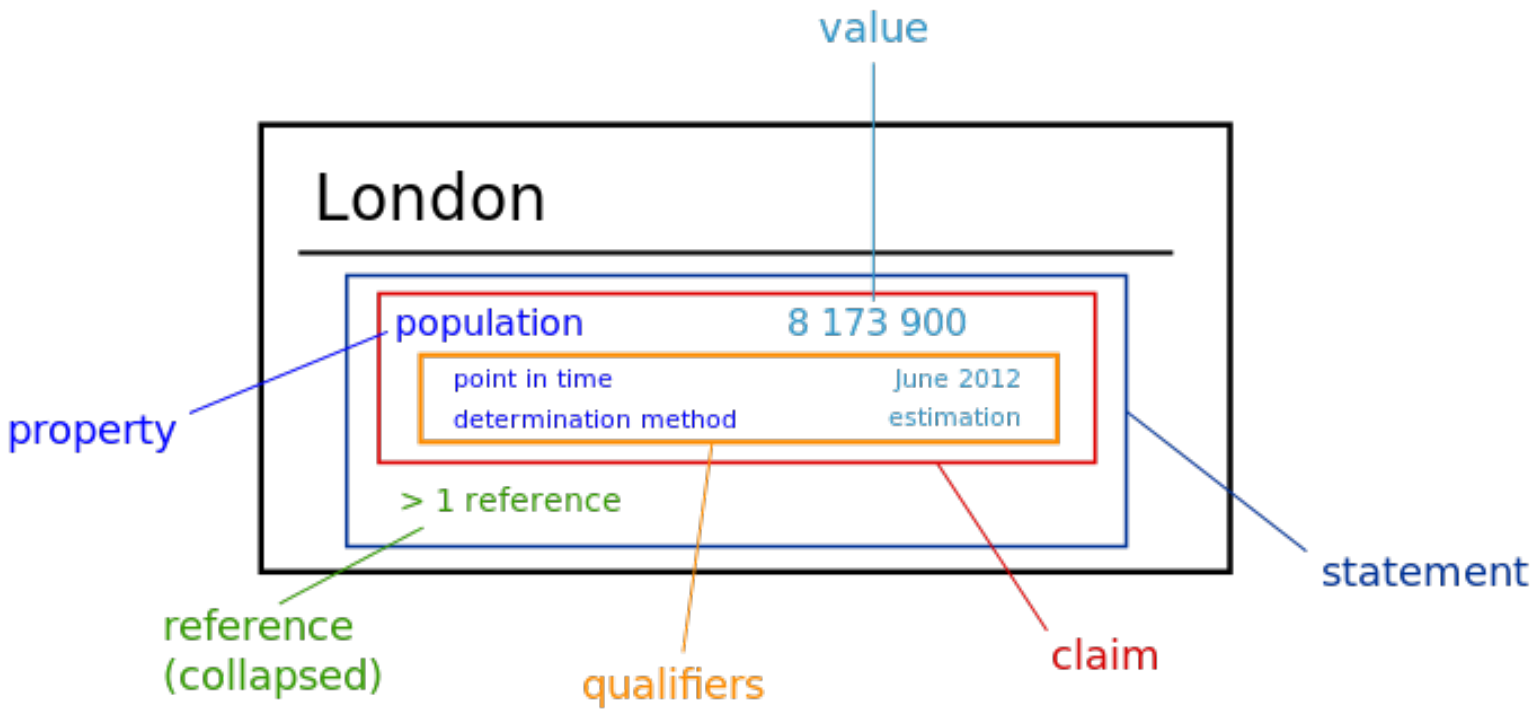


Quelques exemples de KB

Wikidata

En résumé

- tranfert partiel du contenu de Freebase
  - contenu collaboratif
  - import d'autres sources comme MusicBrainz
- plus de 20 millions d'items (nov 2016)



Quelques exemples de KB

YAGO

En résumé

- ontologie avec système de typage très riche
- construit à partie de WordNet de Wikipédia, et en particulier des catégories
- fondé sur une extension de RDFS
- tout est entité :
  - objets : villes, personnes, URL, nombres, dates, mots...
  - classes (hiérarchie)
  - relations
  - faits = (entité, relation, entité)
    - le fait que ce soit une entité permet de donner par exemple sa source
- relations n-aires = un fait principal + d'autres arguments en relation avec ce fait

Objectif du cours

Extraction d'information/Acquisition de connaissances

- sous l'angle du Traitement Automatique des Langues

Composantes principales

- De quoi parle-t-on ?

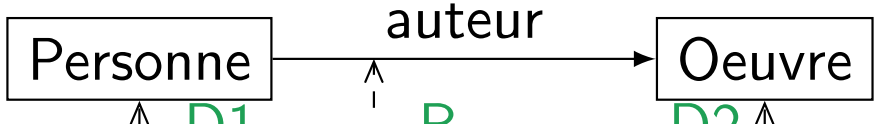
→ entités : qui ? quoi ? quand ? où ?

- Qu'en dit-on ?

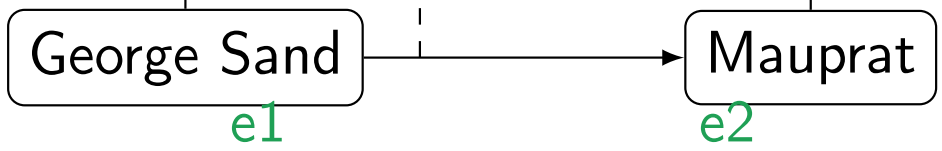
→ relations entre les entités

Quelques définitions

relation et types



entités et instance de relation



mentions

Mauprat$_{m1}$, que Sand$_{m2}$ écrivit$_{mr}$ entre 1835 et 1837, est bien un roman capital dans son œuvre.

Un exemple

Objectif

Trouver la date de début de carrière de Cecilia Bartoli pour la stocker dans une base de connaissances (relation DBPedia)

> En 1985 — elle n'a que 19 ans —, Cecilia Bartoli se fait connaître en France.

Étapes

En 1985$_{DATE}$ — elle n'a que 19 ans —, Cecilia Bartoli$_{PERS}$ se fait connaître en France.

Cecilia Bartoli

1985

Étapes

En 1985$_{DATE}$ — elle n'a que 19 ans —, Cecilia Bartoli$_{PERS}$ se fait connaître en France.

Cecilia Bartoli

1985

Étapes

En 1985$_{DATE}$ — elle n'a que 19 ans —, Cecilia Bartoli$_{PERS}$ se fait connaître en France.

Étapes

En 1985$_{DATE}$ — elle n'a que 19 ans —, Cecilia Bartoli$_{PERS}$ se fait connaître en France.

1. Introduction

2. Entités
   - Définitions
   - Reconnaissance d'EN
   - Désambiguïsation

3. Relations
   - Définitions
   - Extraction de relations
   - Méthodes supervisées
   - Méthodes peu supervisées

4. Conclusion

Entité nommée

Définition(s)

- expression linguistique renvoyant à un référent unique au sein d'une catégorie en contexte
- typiquement : personnes, organisations, lieux
  - entités numériques souvent associées : dates, montant, vitesse...

Exemple de texte annoté

Le 27 avril 2006$_{DATE}$ à Washington$_{LIEU}$, George Clooney$_{PERS}$ et Barack Obama$_{PERS}$ assistent à une conférence de presse sur le Darfour$_{LIEU}$.

Formats d'annotation

- parenthésé
  - [$_{ORG}$ U.N.] official [$_{PERS}$ *Ekeus*] heads for [$_{LOC}$ *Baghdad*] .
- XML
  - <org>U.N.</org> official <personne>Ekeus</personne> heads for <lieu>Baghdad</lieu>.
  - <enamex type="organisation"> U.N.</enamex> official <enamex type="person">Ekeus</enamex> heads for <enamex type="organisation">Baghdad</enamex> . (MUC)
- BIO ou variantes (ex : BILOU=BIO+Last+Unique)

U.N.	NNP	B-NP	B-ORG
official	NN	I-NP	O
Rolf	NNP	I-NP	B-PER
Ekeus	NNP	I-NP	I-PERS
heads	VBZ	I-VP	O
for	IN	B-PP	O
Baghdad	NNP	I-PP	B-LOC
.	.	.	O

Reconnaissance d'EN

Exemple de texte

(...) et Obama assistent à (...)

Reconnaissance d'EN

- reconnaissance
  - identification

Exemple de texte

(...) et Obama assistent à (...)

Reconnaissance d'EN

- reconnaissance
  - identification
  - catégorisation

Exemple de texte

(...) et <personne> Obama </personne> assistent à (...)

Reconnaissance d'EN

- reconnaissance
  - identification
  - catégorisation

- désambiguïsation (*entity linking*)

Exemple de texte

(...) et <personne ref="Barack_Obama"> Obama </personne> assistent à (...)

Définition de la tâche : quelles catégories ?

Quelles catégories ?

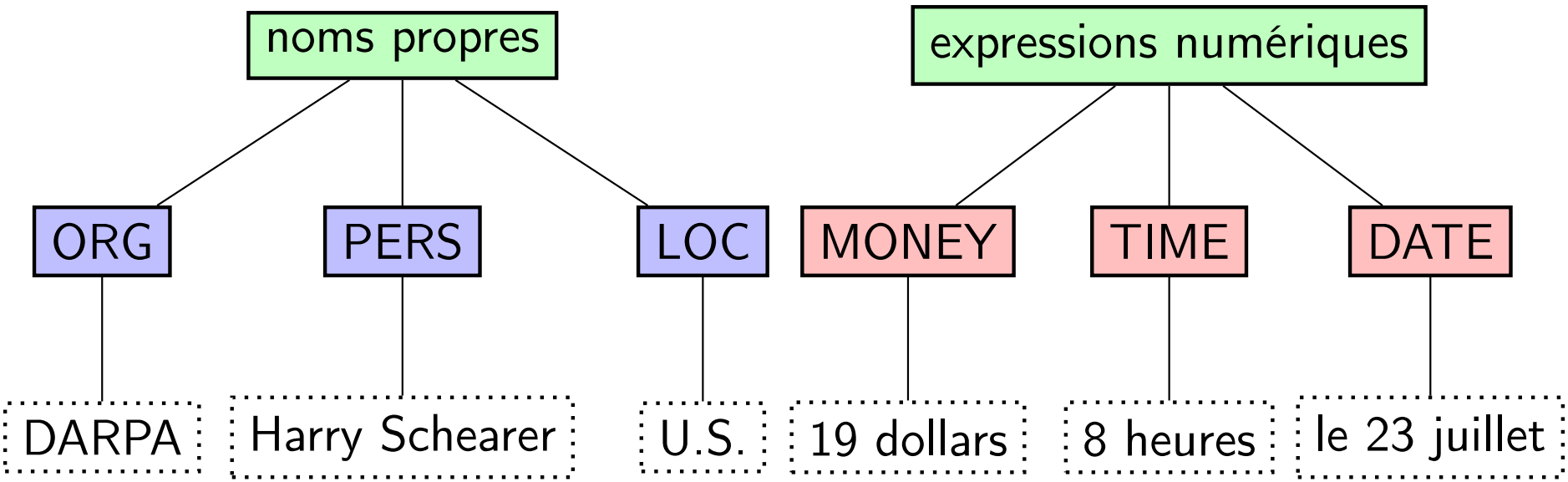
- absence de consensus au-delà des 3 catégories classiques
  - catégorie Divers dans certaines campagnes (CoNLL, HAREM)
- dépendance au type d'application visée
  - granularité des classes : longueur $\neq$ hauteur
- référence à des jeux existants (campagnes d'évaluation)

Portée des catégories ?

- quelles instances ?
  - ☺ Matteo Renzi, la famille Kennedy
  - 😑 Zorro, Hercule, les italiens
  - 😕 Mickey, Bison futé, le Prince Charmant
- ambiguïté, notamment métonymie
  - la France_{ORG} vote contre un traité d'interdiction des armes nucléaires (ou $\textit{France}_{LIEU}$?)

Catégories d'EN

MUC-6/7



Catégories d'EN

ACE (2002-2008)

Caractéristiques

- ajout d'entités non nommées : mentions nominales ou pronominales
- 7 types, dont Person, Organization, Location et :
  - Geo-political Entity
    - France_{ORG} signed a treaty with Germany last week.
    - The world leaders met in France_{LIEU} yesterday.
    - France_{GPE} produces better wine than New Jersey.
  - Facility (*Aéroport Charles de Gaulle*)
  - Vehicle (les *hélicoptères militaires* ont...)
  - Weapon (des *missiles sol-air* ont été tirés)
- hiérarchie : sous-types

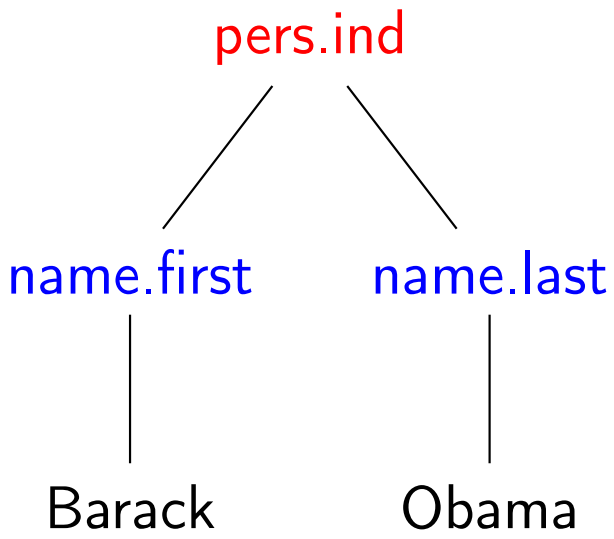
  par exemple pour Person : Individual, Group et Indeterminate (si le contexte ne permet pas de distinguer)

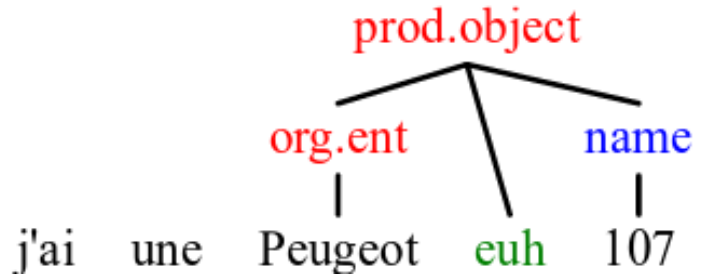
Catégories d'EN

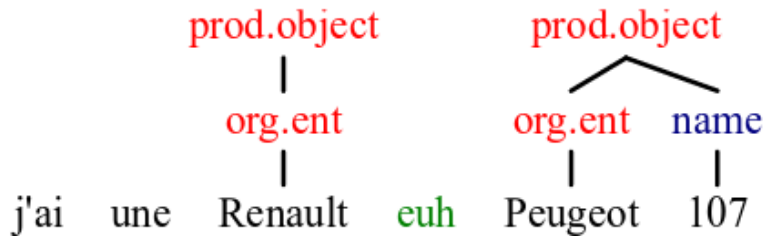
Quaero (2011/2012)

Caractéristiques

- ajout de nouveaux types : produits, fonctions
- structuration supplémentaire : composition
  - prise en compte de métonymie : deux niveaux d'annotation
- annotation adaptée aux corpus oraux (disfluences)

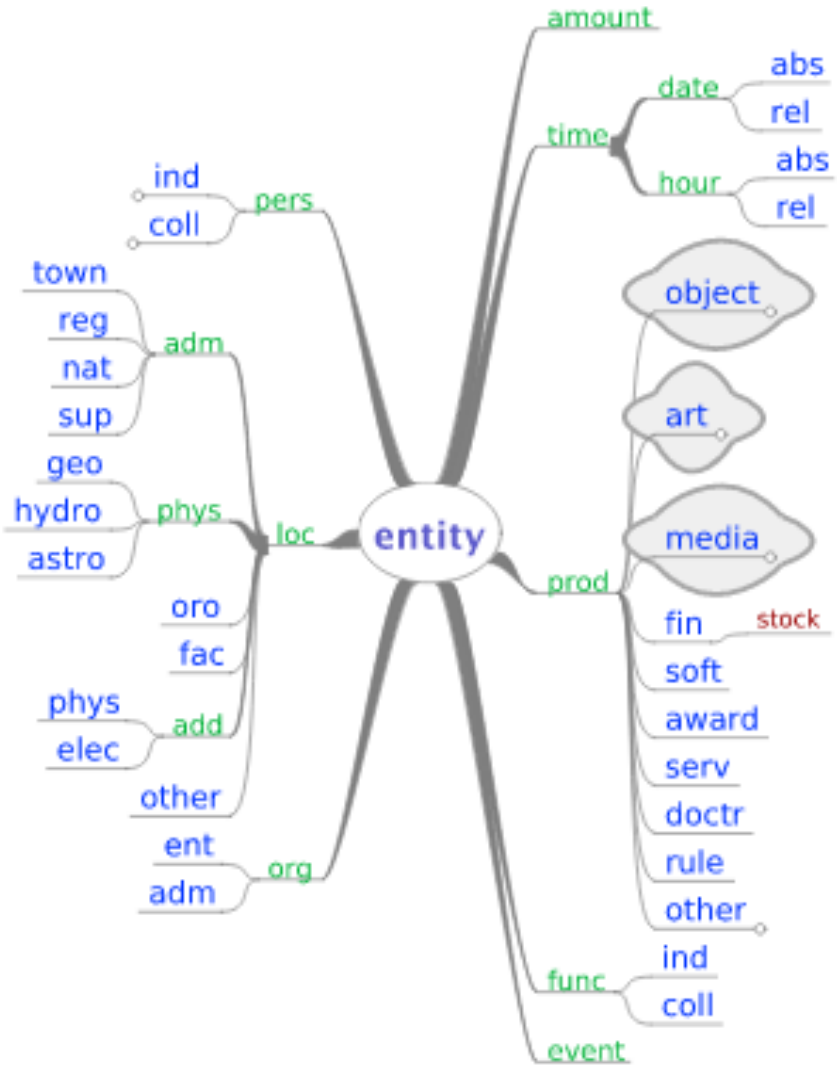




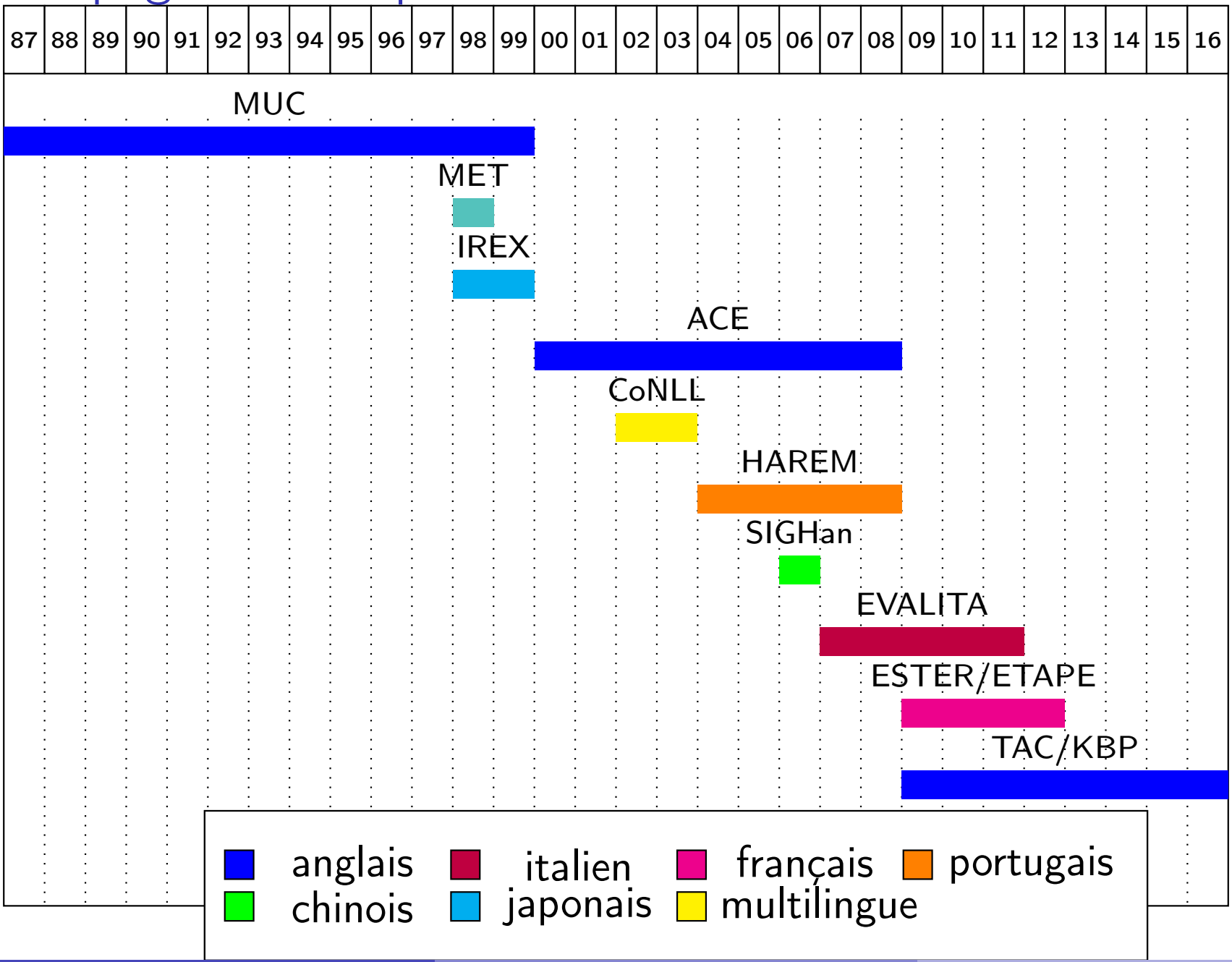


Catégories d'EN

Quaero (2011/2012)



Campagnes et corpus



Définition de la tâche : quelles mentions ?

Portée des annotations

- forme des mentions
  - ☺ noms propres : *Jacques Chirac*
  - ☹ surnoms, groupes nominaux, pronoms : *Chichi*, *l'ancien président*, *il*
- frontières
  - déterminants : *les Rolling Stones*, *La Mecque*, *Le téléphone sonne*
  - fonctions : *le président Obama*, *l'Abbé Pierre*
  - titres : *Monsieur Fillon*, *Professeur Paolucci*
  - désignateurs générationnels : *Benoît XVI*, *Bush Jr.*
- coordination
  - Bill and Hillary Clinton flew to Chicago last month. (ellipse partielle)
  - M. et Mme. Chirac en thalasso à Biarritz. (ellipse totale)

  → Bill and Hillary Clinton$_{PERS}$ vs Bill$_{PERS}$ and Hillary Clinton$_{PERS}$
- imbrication
  - *Université Lyon 2*, *Comité Exécutif d'Orange*

  → Université Lyon 2$_{ORG}$ vs Université Lyon$_{LIEU}$ 2$_{ORG}$ (entité structurée)

Reconnaissance d'EN

Définition

Identifier et catégoriser automatiquement des entités nommées dans des textes

Exemples de difficultés

- homonymie (même type ou type différent)

→ JFK : personne(s) ou aéroport ?, Paris

- métonymie

→ Washington, l'Élysée : lieu (ville) ou organisation ?

Indices pour la reconnaissance ?

Laurent Courtois-Courret, délégué syndical SUD au centre Qualipel, à Vitry-sur-Seine (Val-de-Marne), a écopé de dix jours de mise à pied disciplinaire avec retenue de salaire.

Indices internes

- casse
  - *mRNA* = xXXX, *CPA1* = XXXd
- *n*-grams de caractères :
  - *Cotrimoxazole* → médicament, *Leuville-sur-Orge* → lieu
  - *Twilight - Chapitre 3 : hésitation* → film
- mots classifiant
  - la Banque Populaire
  - l'avenue des Champs-Élysées
  - Benoît XVI (marqueur générationnel)
- sigles ou esperluette
  - Crédit Agricole SA
  - Standard & Poor's
  - F. Hollande
- lexiques (par exemple prénoms), clusters de mots, plongements lexicaux (*word embeddings*)
  - François Hollande

Indices externes

- contexte d'apparition des entités nommées
- informations supplémentaires ou propriétés spécifiques
  - Monsieur Hollande
  - Mme Michel
  - Général Leclerc
  - le groupe Sanofi
  - the Coca Cola company
- souvent précisés uniquement pour la première occurrence de l'entité

Systèmes symboliques

Composants standards

- Reconnaissance de déclencheurs et entités à partir de listes
- Expressions régulières en cascade

Exemple de règle

Université + *de* + NomDeVille ⇒ Organisation

Exemple d'entité reconnue par cette règle

Université de Nantes

Limites

- mauvais rappel : listes incomplètes, évolutivité, entités partielles (*Obama*), textes bruités...
- ambiguïtés (homonymie et métonymie)

Systèmes par apprentissage supervisé

Reformulation en tâche de classification

- entraînement
  - rassembler un corpus représentatif
    - besoin de beaucoup d'exemples annotés !
  - annoter chaque token
  - choisir des attributs adaptés aux classes et aux textes
  - entraîner un classifieur à prédire les étiquettes des tokens
- test
  - annoter chaque token
  - évaluer

token	maj	ponct	prenom	pos	chunk	tag
U.N.	1	1	0	NNP	B-NP	B-ORG
official	0	0	0	NN	I-NP	O
Rolf	1	0	1	NNP	I-NP	B-PER
Ekeus	1	0	0	NNP	I-NP	I-PERS
heads	0	0	0	VBZ	I-VP	O
for	0	0	0	IN	B-PP	O
Baghdad	1	0	0	NNP	I-PP	B-LOC

Systèmes par apprentissage supervisé

Attributs standards

- Mots
  - courant
  - sous-chaînes du mot
  - précédent
  - suivant
- Autres informations linguistiques apprises
  - catégories morpho-syntaxiques

Modèles d'annotation

- étiquettes indépendantes peu adapté
- annotation de suite d'étiquettes avec un sens de parcours
  - limites : fenêtre fixée, propagation des erreurs
- annotation de séquences (CRF)

Approches actuelles

Objectif

Se passer de connaissances a priori et de sélection des attributs

- réseaux de neurones profonds [Collobert et al., 2011]

Résultats récents

- [Lample et al., 2016] : LSTM-CRF, aucune donnée externe
- [Guo et al., 2014, Passos et al., 2014] : CRF + word embeddings
- F1 $\simeq$ 0,90 sur données CoNLL 2003 pour l'anglais (PER, LOC, ORG, MISC)

Évaluation

- tp (vrais positifs) = entités correctement reconnues
- fp (faux positifs) = entités reconnues à tort
- fn (faux négatifs) = entités non reconnues

Métriques classiques

- Précision = $\frac{tp}{tp+fp}$

→ entités correctement annotées par rapport à l'ensemble des entités annotées par le système

- Rappel = $\frac{tp}{tp+fn}$

→ entités correctement annotées par rapport à l'ensemble des entités qu'il fallait annoter

Évaluation

Référence

<personne>Jean-Yves Le Drian</personne> engage ses homologues à "parler d'abord de manière européenne" sur le plan militaire.

Hypothèse (sortie du système)

<personne>Jean-Yves</personne> Le Drian engage ses homologues à "parler d'abord de manière européenne" sur le plan militaire.

Inconvénient de ces métriques pour les entités nommées

- *Jean-Yves* reconnu comme une entité à tort

→ faux positif

- *Jean-Yves Le Drian* non reconnu

→ faux négatif

Métriques adaptées

- R : # entités dans la référence
- H : # entités dans l'hypothèse (= sortie du système)
- C : # entités correctes (= vrais positifs)
- T : # entités avec les bonnes frontières mais mal catégorisées
- F : # entités bien catégorisées mais avec les mauvaises frontières
- TF : # entités avec mauvais type et frontières
- I : # entités insérées (= faux positifs)
- D : # entités oubliées (= faux négatifs)

Métriques adaptées

→ reconnaissance partielle comptée comme à moitié bonne

- Précision = $\frac{C+0.5\times(T+F)}{H}$
- Rappel = $\frac{C+0.5\times(T+F)}{R}$
- Slot Error Rate = $\frac{D+I+TF+0.5\times(T+F)}{R}$

Exemple d'évaluation

Référence (annotation manuelle)

<personne>Bertrand Delanoë</personne> a été élu maire de <lieu>Paris</lieu>.

Hypothèse 1 (système 1)

<personne>Bertrand Delanoë</personne> a été élu maire de <personne>Paris</personne>.

SER = (0 + 0 + 0 + 0,5 * (1 + 0)) / 2 = 0,25

Hypothèse 2 (système 2)

<personne>Bertrand</personne> Delanoë a été élu maire de <personne>Paris</personne>.

SER = (0 + 0 + 0 + 0,5 * (1 + 1)) /2 = 0,5

Mon exemple

En 1985$_{DATE}$ — elle n'a que 19 ans —, Cecilia Bartoli$_{PERS}$ se fait connaître en France.

Cecilia Bartoli

1985

Désambiguïsation - définition

Désambiguïsation/résolution/liaison (*entity linking*)

Étant donné une base de connaissances, choisir l'entité correspondant à la mention (référent)

Texte à analyser

In a grim preview of the discontent that may cloud at least the outset of the next president's term, Hillary Clinton and Donald J. Trump are seen by a majority of voters as unlikely to bring the country back together after this bitter election season.
With more than eight in 10 voters saying the campaign has left them repulsed rather than excited, the rising toxicity threatens the ultimate victor. Mrs. Clinton, the Democratic candidate, and Mr. Trump, the Republican nominee, are seen as dishonest and viewed unfavorably by a majority of voters.

Désambiguïsation - définition

Désambiguïsation/résolution/liaison (*entity linking*)

Étant donné une base de connaissances, choisir l'entité correspondant à la mention (référent)

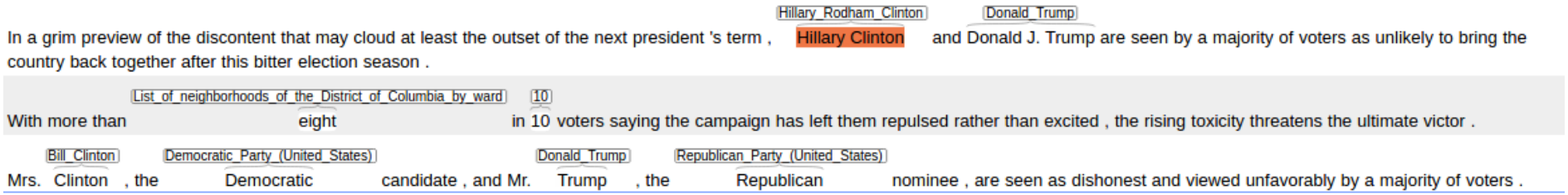
Résultat attendu

In a grim preview of the discontent that may cloud at least the outset of the next president's term, Hillary Clinton$_{Hillary_Clinton}$ and Donald J. Trump$_{Donald_Trump}$ are seen by a majority of voters as unlikely to bring the country back together after this bitter election season.
With more than eight in 10 voters saying the campaign has left them repulsed rather than excited, the rising toxicity threatens the ultimate victor. Mrs. Clinton$_{Hillary_Clinton}$, the Democratic$_{Democratic_Party_(United_States)}$ candidate, and Mr. Trump$_{Donald_Trump}$, the Republican$_{Republican_Party_(United_States)}$ nominee, are seen as dishonest and viewed unfavorably by a majority of voters.

Désambiguïsation - définition

Désambiguïsation/résolution/liaison (*entity linking*)

Étant donné une base de connaissances, choisir l'entité correspondant à la mention (référent)



CoreNLP

Désambiguïsation - définition

Désambiguïsation/résolution/liaison (*entity linking*)

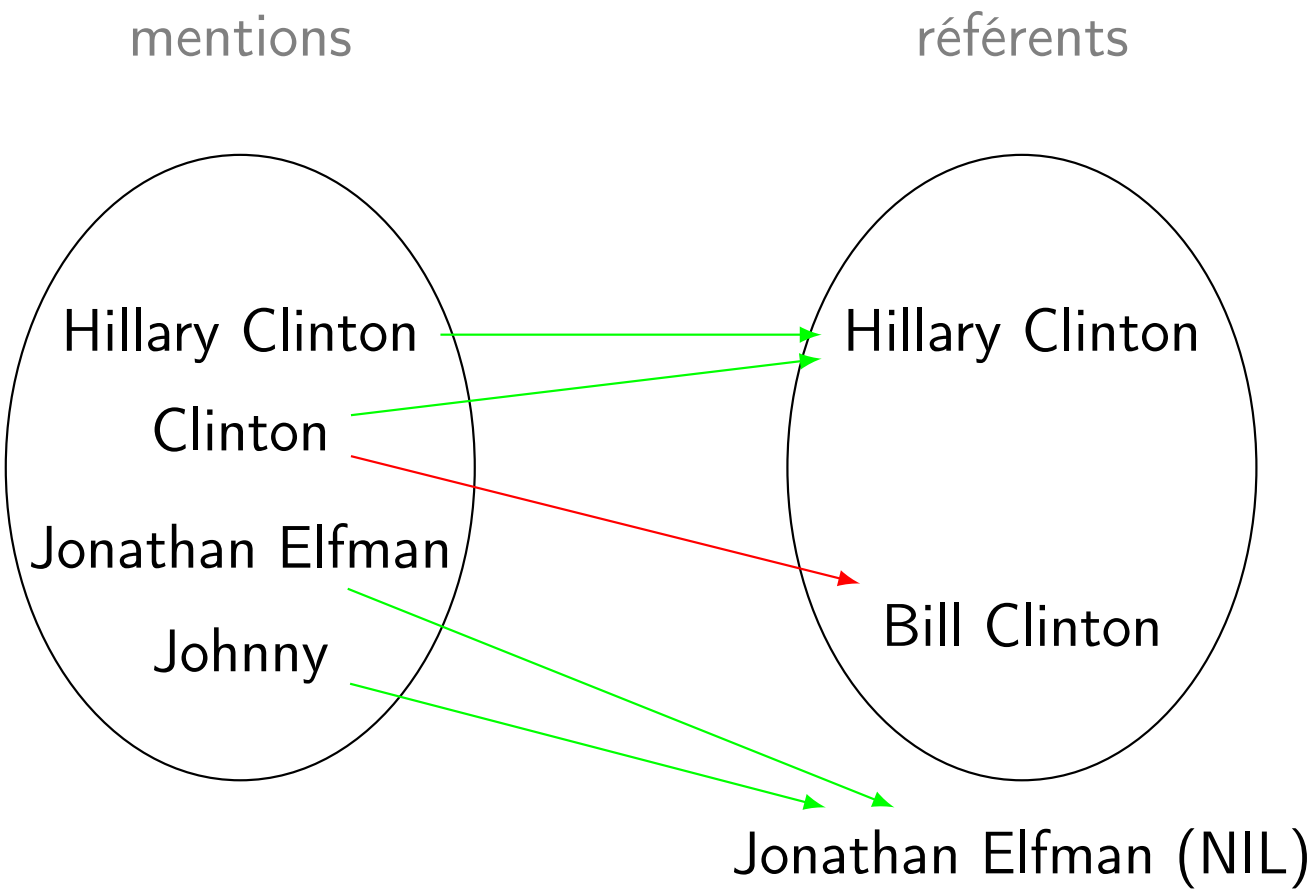
Étant donné une base de connaissances, choisir l'entité correspondant à la mention (référent)

In a grim preview of the discontent that may cloud at least the outset of the next president's term, Hillary Clinton and Donald J. Trump are seen by a majority of voters as unlikely to bring the country back together after this bitter election season.

With more than eight in 10 voters saying the campaign has left them repulsed rather than excited, the rising toxicity threatens the ultimate victor. Mrs. Clinton, the Democratic candidate, and Mr. Trump, the Republican nominee, are seen as dishonest and viewed unfavorably by a majority of voters.

DBPedia Spotlight

Désambiguïsation - difficultés



Désambiguïsation - principe

Étapes

1. détection de possibles mentions
   - souvent fondée sur reconnaissance d'EN
2. sélection de candidats
   - proximité graphique avec labels, texte des liens, requêtes menant aux pages Wikipédia, pages Wikipédia de désambiguïsation
3. ordonnancement des candidats

   WSD / Wikipédia
   - mention : distance avec les labels des référents
   - référent : popularité (plus fréquent, page Wikipédia avec le plus de liens...)
   - contexte local de la mention : similarité textuelle avec pages Wikipédia, des liens...
   - contexte global de la mention (document) : autres entités (désambiguïsation collective), coréférence

KBP

Tâche Entity Discovery and Linking

- Discovery : détecter et annoter mentions
  - classes : LOC, ORG, PER, FAC, GPE ;
  - mentions : EN, noms, auteurs de posts
- Linking : rattacher clusters de mentions à une base de connaissances
- Difficultés (KBP 2015) :
  - détection des noms communs et abréviations
  - entités rares
  - biais de popularité
  - connaissances générales
  - langue informelle
  - incohérence entre type EN et référent
- F1 $\simeq$ 0,60 pour EL en anglais 2015, 0,80 en 2014

Mon exemple

En 1985$_{DATE}$ — elle n'a que 19 ans —, Cecilia Bartoli$_{PERS}$ se fait connaître en France.

Cecilia Bartoli

1985

1. Introduction

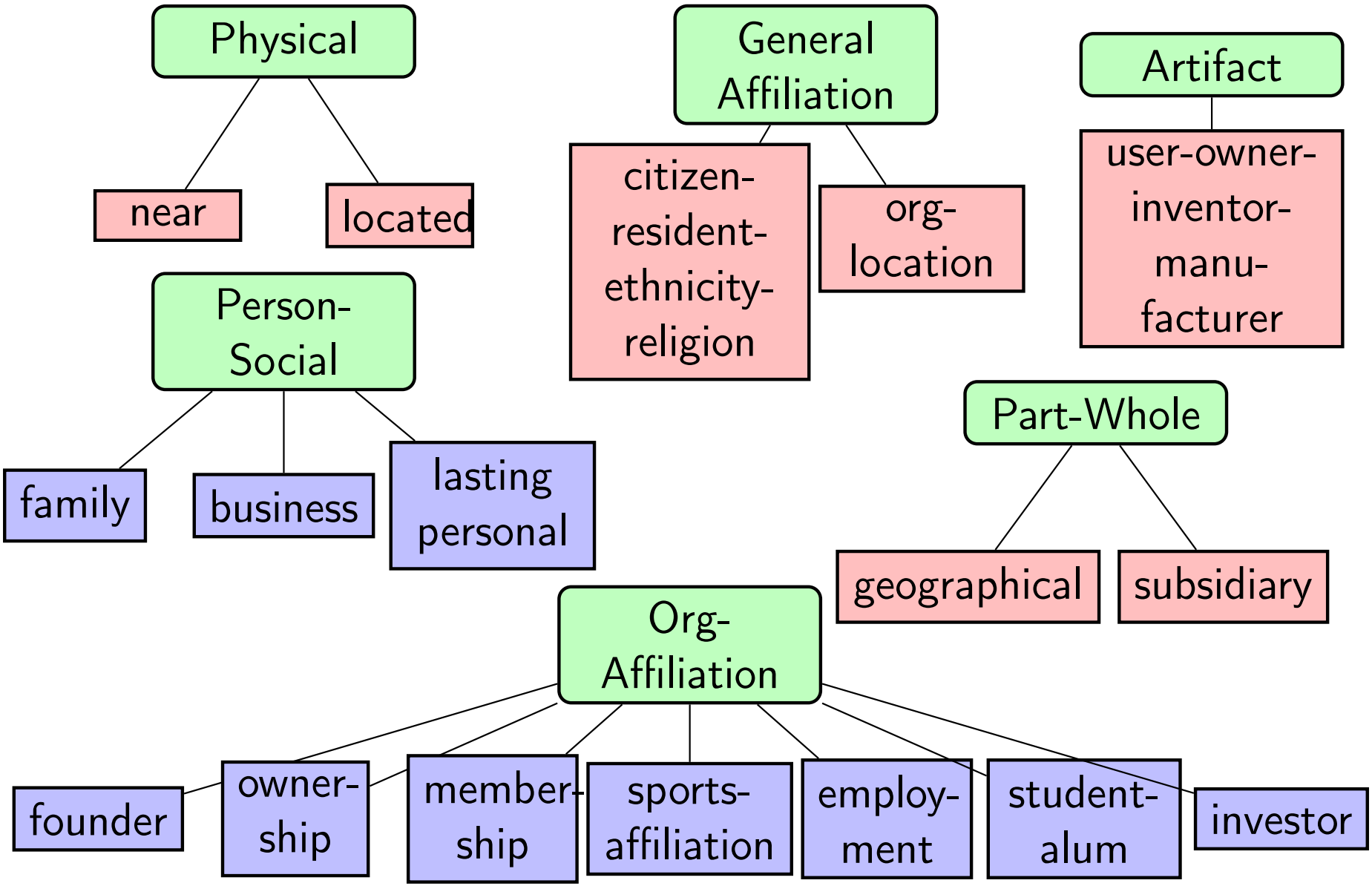
2. Entités
   - Définitions
   - Reconnaissance d'EN
   - Désambiguïsation

3. Relations
   - Définitions
   - Extraction de relations
   - Méthodes supervisées
   - Méthodes peu supervisées

4. Conclusion

Quelques exemples de jeux de relations

ACE 2005



Quelques exemples de jeux de relations

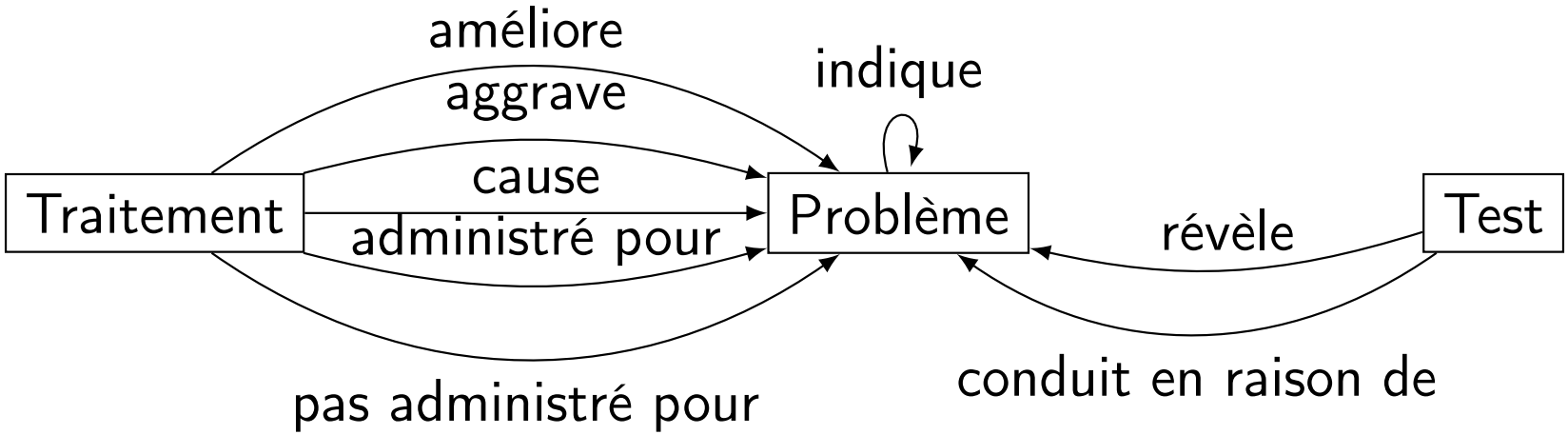
SemEval 2010 tâche 8

Type	Exemple
Cause-Effect	The *news* brought about a *commotion* in the office.
Instrument-Agency	*Carpenters* build many things from *wood*.
Product-Producer	The *government* built 10,000 new *homes*.
Content-Container	I emptied the *wine bottle* into my glass.
Entity-Origin	It involves a spectator choosing a *card* from the *deck*.
Entity-Destination	He sent his *painting* to an *exhibition*.
Component-Whole	Feel free to download the first *chapter* of the *book*.
Member-Collection	A person who is serving on a *jury* is known as *juror*.
Message-Topic	Mr Cameron asked a *question* about tougher *sentences* for people carrying knives.

Quelques exemples de jeux de relations

i2b2 2010

Analyse de comptes rendus cliniques



Quelques exemples de jeux de relation

Freebase

Relations Freebase les plus fréquentes

- /people/person/nationality
- /location/location/contains
- /people/person/location
- /people/person/place_of_birth
- /dining/restaurant/cuisine
- /business/business_chain/location
- /biology/organism_classification_rank
- /film/film/genre
- /film/film/language
- /biology/organism_higher_classification
- /film/film/country
- /film/writer/film

Relation

Caractéristiques des relations

- entre concepts ou entre instances de concepts
- hiérarchiques ou autres
- prise en compte des événements ou ensemble de relations binaires
- relations "du monde réel" ou avec factivité...

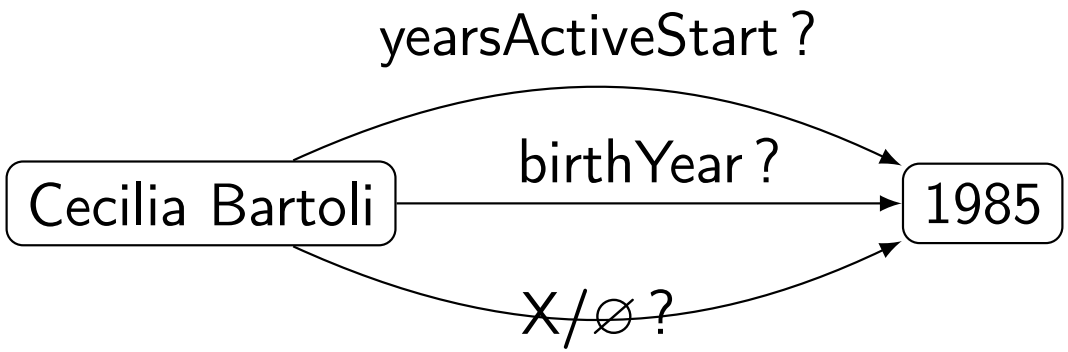
Extraction de relations

Définition

Étant donné deux (ou plus) entités, déterminer

- si elles sont en relation
- le type de relation

En 1985$_{DATE}$ — elle n'a que 19 ans —, Cecilia Bartoli$_{PERS}$ se fait connaître en France.



Difficultés

Variabilité d'expression des relations

- En 1985 — elle n'a que 19 ans —, Cecilia Bartoli se fait connaître en France lors d'un concert organisé par l'Opéra de Paris en hommage à Maria Callas.
- C'est déjà une longue carrière que celle de Cecilia Bartoli. Elle débute en 1985, à Rome. Elle a dix-neuf ans et incarne la pétulante Rosina du « Barbier de Séville ».
- En 1985, une tournée en Allemagne de l'Est et un gala télévisé à Paris en hommage à Maria Callas suffisent à attirer l'attention de tous – y compris celle de chefs d'orchestre prestigieux comme Daniel Barenboim, Claudio Abbado, Simon Rattle, Herbert von Karajan – sur cette jeune cantatrice.

exemples wikipédia, les échos et encyclopédie universalis

Méthodes simples

Cooccurrence

mais ambiguïté

- personne - date : date de début de carrière, de naissance, autre ?
- traitement - maladie : guérit ? prévient ? effet secondaire ?

Patrons lexico-syntaxiques

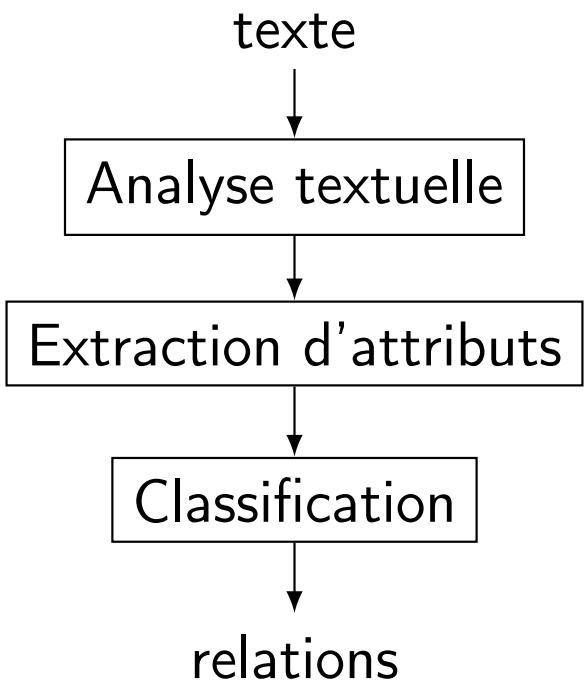
- par exemple pour relation birthyear :
  - *PERSON*, né en *DATE*
  - *PERSON* (*DATE*-)
  - *PERSON* est un GN né en *DATE*
- patrons à écrire manuellement pour chaque relation
  - acquisition automatique possible cependant
  - bootstrapping
- orientés rappel ou précision

Extraction supervisée de relations

Formulé comme un problème de classification

- Classification binaire ou multi-classes
- Exemples d'entraînement positifs et négatifs

Méthode supervisée classique



Modélisation du contexte

En **1985**— elle n'a que 19 ans —, **Cecilia Bartoli** se fait connaître en France

Attributs

- mots (ou lemmes)
  - des différentes parties du contexte
  - sacs de mots et n-grams
  - têtes syntaxiques et concaténation
- types d'entités
  - type des entités et concaténation
- informations syntaxiques
  - chemin de constituants
  - chemin de dépendances
- ressources externes
  - listes de pays, déclencheurs...

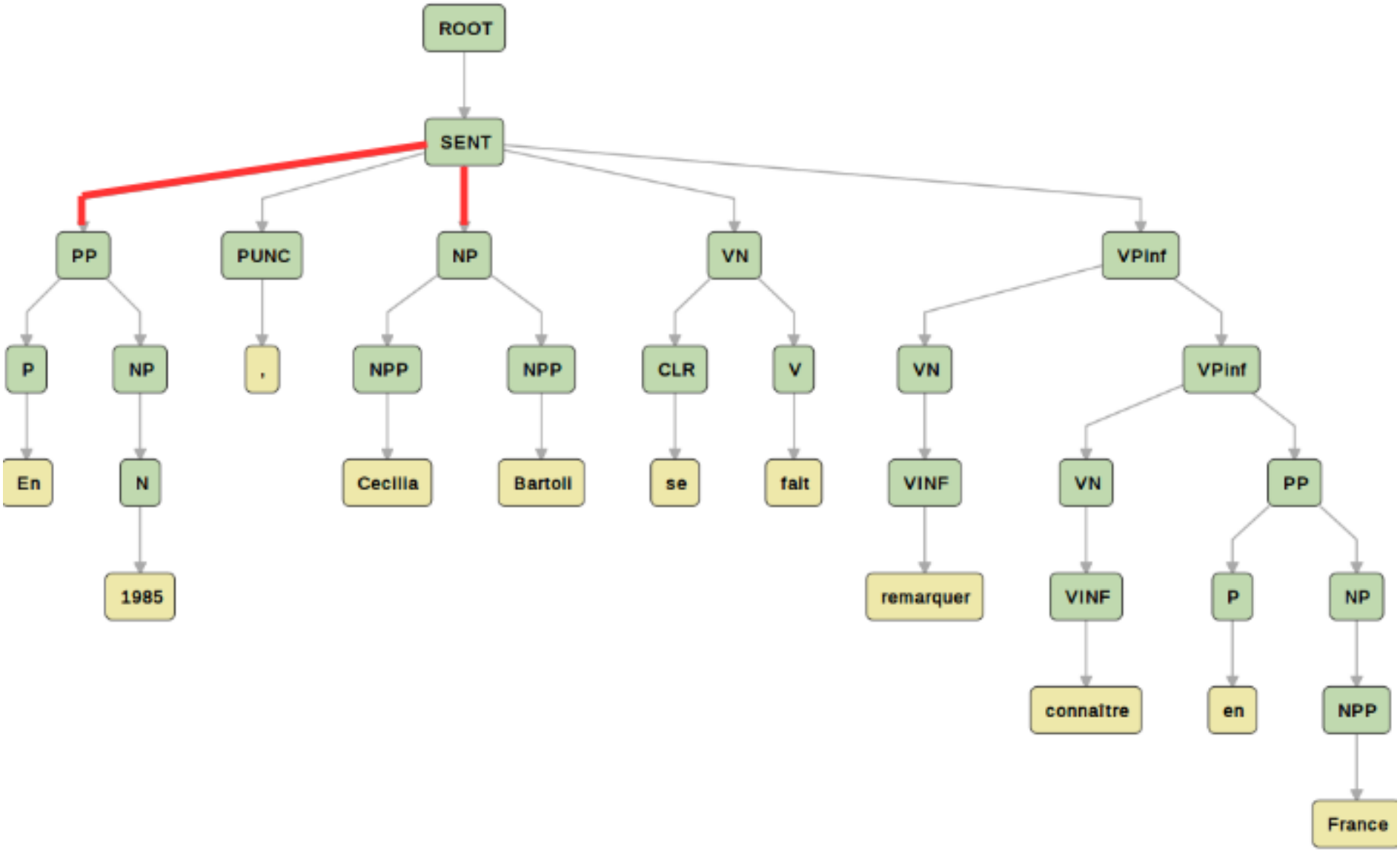
Exemple d'attributs

En **1985**— elle n'a que 19 ans —, **Cecilia Bartoli** se fait connaître en France

- mots
  - avant e_1 : {En}
  - entre les entités (bow) : {elle, n', a, que 19, ans}
  - après e_2 : {se, fait, connaître, en, France}
  - tête e_1 : 1985
  - tête e_2 : Bartoli
- types
  - type e_1 : DATE
  - type e_2 : PERSON
  - concaténation : DATEPERSON
- syntaxe
  - constituants : PP - SENT - NP
  - dépendances : nmod -suj

Exemple

Arbre de constituants



Représentations structurées

Quels attributs ?

- intuition
- expériences

Utiliser les représentations structurées

Définition de métriques de similarité appropriées : noyaux pour arbres syntaxiques

Expériences

- arbre des constituants [Zelenko et al., 2003]
- arbre des dépendances [Culotta and Sorensen, 2004]
- plus court chemin de dépendances entre entités [Bunescu and Mooney, 2005]

Limites des approches classiques

Inconvénients des méthodes précédentes

- qualité de la classification fortement dépendante des prétraitements
- grands corpus annotés
  - même si myriadisation (crowdsourcing) possible [Liu et al., 2016]
- déséquilibre des corpus
- manque de généralisation

S'affranchir des prétraitements

Utilisation de réseaux de neurones profonds

- en entrée : mots, *n*-grams + positions + plongements lexicaux
- réseau : RNN ou CNN

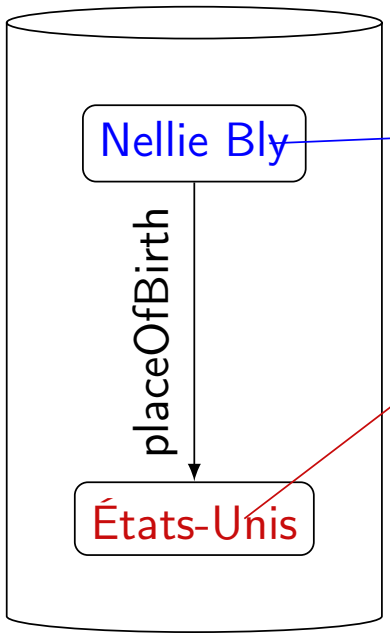
Supervision distante

Objectif

- annoter automatiquement les exemples d'entraînement

$\leftarrow$ bases de connaissances

- puis méthodes classiques



Nellie Bly est née le 5 mai 1864 en Pennsylvanie, aux Etats-Unis.

Figure culte aux Etats-Unis, Nellie Bly fut la première femme à pratiquer le journalisme dit "en immersion".

Nellie Bly arrive aux Etats Unis le 25 janvier 1890 réalisant ainsi, en 72 jours, une performance extraordinaire.

relations dbpedia

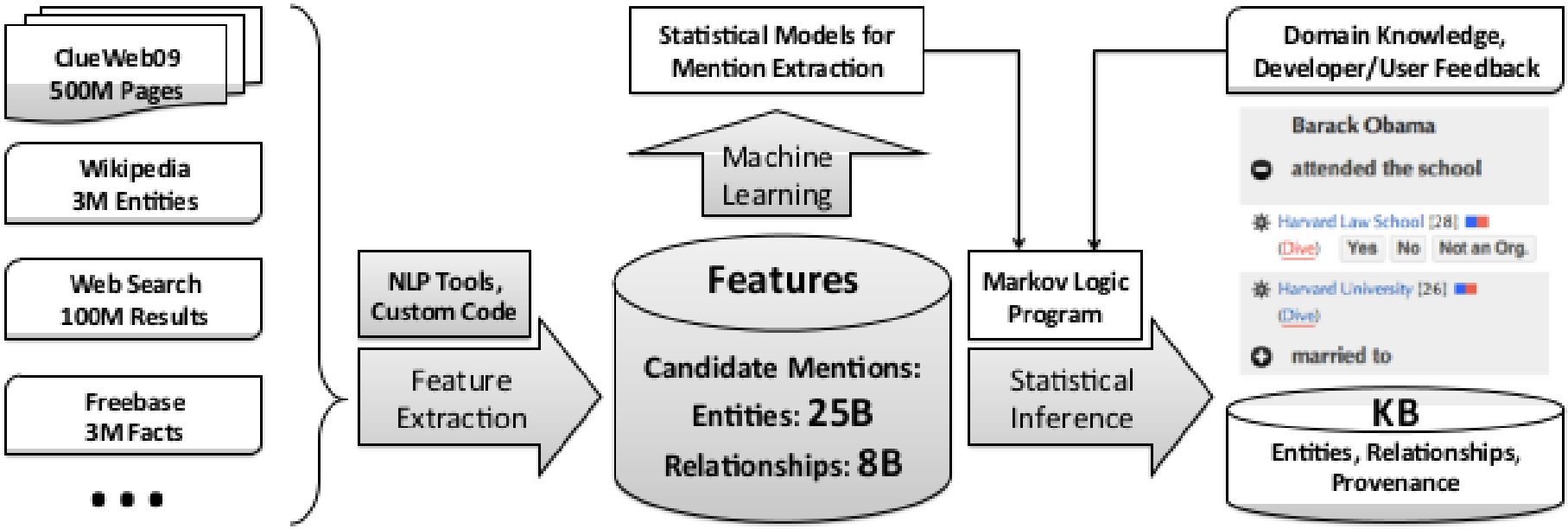
Supervision distante

Hypothèses

- Une phrase contenant les entités participant à une relation donnée a de fortes chances d'exprimer cette relation
  [Mintz et al., 2009, Wu and Weld, 2007, Niu et al., 2012]
- Problème d'apprentissage multi-instances [Riedel et al., 2010] : au moins une des phrases contient une mention de la relation
- Plusieurs relations peuvent exister entre deux entités [Hoffmann et al., 2011]

- ☺ (relativement) indépendant du domaine
- ☺ passage à l'échelle
- ☹ valable uniquement pour les relations hors contexte
- ☹ dépend de la qualité de la reconnaissance d'entité

DeepDive [Niu et al., 2012]



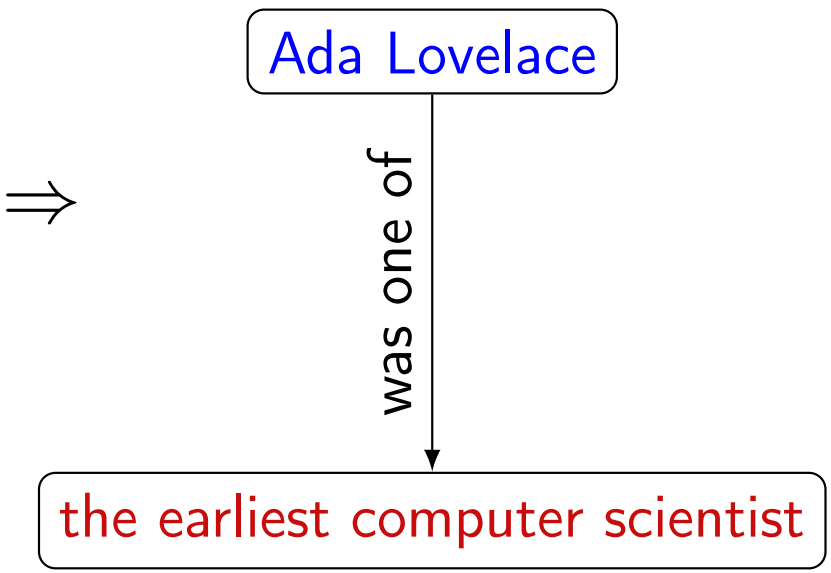
Extraction d'information ouverte

Principe

Partir des expressions de relations dans des textes

Ada Lovelace was one of the earliest computer scientists.

The second tunnel boring machine will be named Ada after Ada Lovelace who was one of the earliest computer scientists.

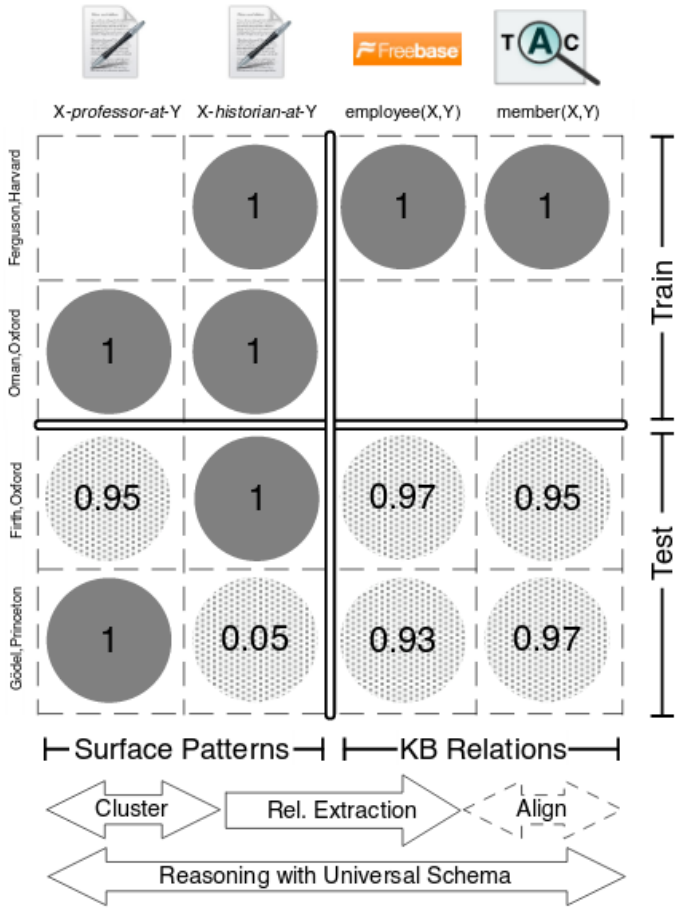


(exemples de `http://openie.allenai.org/`)

Extraction d'information ouverte

Limites

- Relations non normalisées
  - [Angeli et al., 2015] : cooccurrences relations OIE et KBP en corpus
  - [Riedel et al., 2013, Verga et al., 2016] : implications entre relations



Difficultés actuelles

- relations rares (dans les textes)
- relations contextuelles
- factivité
- source : fiabilité, fiction...
- connaissances de sens commun
- interrogation NL : plusieurs relations

Mon exemple

En 1985_{DATE} — elle n'a que 19 ans —, Cecilia Bartoli$_{PERS}$ se fait connaître en France.

Conclusion

Quelques points

- connaissances de plus en plus présentes explicitement
- cercle vertueux entre EI et annotation sémantique
- ressources restent complémentaires
  - interrogation des deux types de ressources
  - réelle interaction entre raisonnement sur textes et connaissances

Documents de référence

- Cours de Christopher Manning et Dan Jurafsky sur le traitement automatique des langues (Natural Language Processing)
- Livre Les entités nommées pour le traitement automatique des langues, Damien Nouvel, Maud Ehrmann et Sophie Rosset, 2015

Références I

Angeli, G., Premkumar, M. J., and Manning, C. D. (2015).
Leveraging linguistic structure for open domain information extraction.
In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.

Bunescu, R. and Mooney, R. (2005).
A shortest path dependency kernel for relation extraction.
In *Proceedings of the conference on human language technology and empirical methods in natural language processing*.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. (2011).
Natural language processing (almost) from scratch.
Journal of Machine Learning Research, 12(Aug) :2493–2537.

Culotta, A. and Sorensen, J. (2004).
Dependency tree kernels for relation extraction.
In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 423. Association for Computational Linguistics.

Guo, J., Che, W., Wang, H., and Liu, T. (2014).
Revisiting embedding features for simple semi-supervised learning.
In *EMNLP*, pages 110–120.

Références II


Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., and Weld, D. S. (2011).
Knowledge-based weak supervision for information extraction of overlapping relations.
In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies-Volume 1*, pages 541–550. Association for Computational Linguistics.


Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., and Dyer, C. (2016).
Neural architectures for named entity recognition.
In *Proceedings of NAACL-HLT 2016*.


Liu, A., Soderland, S., Bragg, J., Lin, C. H., Ling, X., and Weld, D. S. (2016).
Effective Crowd Annotation for Relation Extraction.
In *Proceedings of NAACL-HLT 2016*.


Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009).
Distant supervision for relation extraction without labeled data.
In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP : Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.


Niu, F., Zhang, C., Ré, C., and Shavlik, J. W. (2012).
DeepDive : Web-scale Knowledge-base Construction using Statistical Learning and Inference.
VLDS, 12 :25–28.

Références III


Passos, A., Kumar, V., and McCallum, A. (2014).
Lexicon infused phrase embeddings for named entity resolution.
In *Proceedings of the Eighteenth Conference on Computational Language Learning*.


Riedel, S., Yao, L., and McCallum, A. (2010).
Modeling Relations and Their Mentions without Labeled Text.
In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.


Riedel, S., Yao, L., McCallum, A., and Marlin, B. M. (2013).
Relation extraction with matrix factorization and universal schemas.
In *Proceedings of NAACL-HLT 2013*.


Verga, P., Belanger, D., Strubell, E., Roth, B., and McCallum, A. (2016).
Multilingual relation extraction using compositional universal schema.
In *Proceedings of NAACL-HLT 2016*.


Wu, F. and Weld, D. S. (2007).
Autonomously semantifying wikipedia.
In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 41–50. ACM.


Zelenko, D., Aone, C., and Richardella, A. (2003).
Kernel methods for relation extraction.
Journal of machine learning research, 3(Feb) :1083–1106.