

# Extraction d'information

Anne-Laure Ligozat

2017/2018

## Documents textuels et bases de connaissances

article

tweets

wikipédia

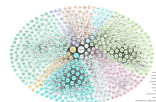
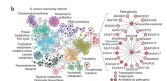
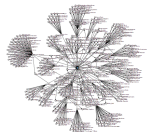
page web

wikidata

wordnet

gene ontology

linking open data



# Documents textuels et connaissances structurées

interrogation et complétion des données structurées

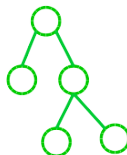
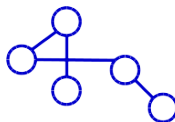
non structuré

A quelques jours d'élections professionnelles qui se dérouleront dans un climat tendu, **SUD** confirme être devenu une cible pour la direction d'**Ad**. Coup sur coup, trois représentants du syndicat ont été visés par des sanctions importantes dans le groupe fondé par **Xavier Niel** (qui détient la marque **Free**). Deux d'entre eux ont fait l'objet d'un entretien préalable au licenciement **vendredi 25 octobre**.

En **ce jeudi après-midi**, malgré une température démente, les cache-nez sont de rigueur aux abords de la **rue Hénard**, dans le **12e arrondissement de Paris**. 2 à 300 policiers « en colère » se massent devant un cordon de sécurité dressé à proximité des locaux de l'**IGPN** (Inspection générale de la police nationale). Ils sont venus soutenir leur collègue **Guillaume Lebeau**, agent de la **BAC** des **Hauts-de-Seine**, auditionné pour s'être répandu à visage découvert dans les médias lors des précédentes manifestations spontanées.

Au cours d'un deuxième débat tendu, l'ex-chef de l'Etat a servi de punching-ball à ses adversaires. Incarnation présidentielle, alliances diplomatiques douteuses, inconstance politique... Les candidats ont décidé d'user de leur droit d'inventaire sur les années **Sarkozy**.

(exemples Mediapart)



structuré

ajout de structure au texte

# 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

## Résumé compréhensible par un programme



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" ?



**Lawrence Livermore National Laboratory**

From Wikipedia, the free encyclopedia

The **Lawrence Livermore National Laboratory (LLNL)** in Livermore, California is a scientific research laboratory founded by the University of California in 1952. It is funded by the United States Department of Energy (DOE) and managed by Lawrence Livermore National Security, LLC (LLNS), a partnership of the University of California, Bechtel Corporation, Babcock and Wilcox, the URS Corporation, and Battelle Memorial Institute. On October 1, 2007 LLNS assumed management of LLNL from the University of California, which had exclusively managed and operated the Laboratory since its inception 55 years before.

**Contents** (hide)

- 1 Background
- 2 Origins
- 3 Weapons projects
- 4 Plutonium research
- 5 National Ignition Facility and proton science
- 6 Global security program
- 7 Other programs
- 8 Key accomplishments
- 9 Unmanned facilities
- 10 Workstations, computers
- 11 Spinoffs
- 12 Directors
- 13 Organization
- 14 Facilities
- 15 References
- 16 External links and sources

**Background**

LLNL is self-described as "a premier research and development institution for science and technology applied to national security."<sup>1</sup> Its principal responsibility is ensuring the safety, security and reliability of the nation's nuclear weapons through the application of advanced science, engineering and technology. The Laboratory also applies its special expertise and multidisciplinary capabilities to preventing the proliferation and use of weapons of mass destruction, bolstering homeland security and solving other nationally important problems, including energy and environmental security, basic science and economic competitiveness.

LLNL is home to many unique facilities and a number of the most powerful computer systems in the world, according to the TOP500 list, including Blue Gene/L, the world's fastest computer from 2004 until Los Alamos National Laboratory's Roadrunner supercomputer surpassed it in 2008. The Lab is a leader in technical innovation: since 1978, LLNL has received a total of 118 prestigious R&D 100 Awards, including

"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

Google  🔍

Tous Maps Actualités Images Vidéos Plus ▼ Outils de recherche

---



**Lyon / Lieux d'intérêt**

<b>Vieux Lyon</b> Renaissance, histoire, jardin		<b>Primatiale Saint-Jean de Lyon</b> Cathédrale	
<b>Basilique Notre-Dame de Fourvière</b> Église à plan basilical		<b>Parc de la Tête d'Or</b> Parc	
<b>Fourvière</b> Musée, théâtre, amphithéâtre, cathédrale, église		<b>Musée des beaux-arts de Lyon</b> Musée, musée d'art, art, architecture	

---

**Office du Tourisme de Lyon: Accueil**  
[www.lyon-france.com/](http://www.lyon-france.com/) ▼  
Only Lyon - Tourisme et Congrès ..... Sandrine, conseiller séjour, vous accueille aujourd'hui lundi 7 novembre, au pavillon du tourisme place Bellecour. Lyon ...

**L'Office du Tourisme - Office du Tourisme de Lyon**  
[www.lyon-france.com/L-Office-du-Tourisme](http://www.lyon-france.com/L-Office-du-Tourisme) ▼











# Applications de l'EI de la vie courante

Google  🔍

Tous Maps Actualités Images Vidéos Plus ▼ Outils de recherche

---



**Lyon** / Lieux d'intérêt

<b>Vieux Lyon</b> Renaissance, histoire, jardin		<b>Primatiale Saint-Jean de Lyon</b> Cathédrale	
<b>Basilique Notre-Dame de Fourvière</b> Église à plan basilical		<b>Parc de la Tête d'Or</b> Parc	
<b>Fourvière</b> Musée, théâtre, amphithéâtre, cathédrale, église		<b>Musée des beaux-arts de Lyon</b> Musée, musée d'art, art, architecture	

---

**Office du Tourisme de Lyon: Accueil**  
[www.lyon-france.com/](http://www.lyon-france.com/) ▼  
Only Lyon - Tourisme et Congrès ..... Sandrine, conseiller séjour, vous accueille aujourd'hui lundi 7 novembre, au pavillon du tourisme place Bellecour. Lyon ...

**L'Office du Tourisme - Office du Tourisme de Lyon**  
[www.lyon-france.com/L-Office-du-Tourisme](http://www.lyon-france.com/L-Office-du-Tourisme) ▼



# Applications de l'EI de la vie courante

Google limsi adresse


Tous Maps Actualités Images Shopping Plus ▾ Outils de recherche

Environ 10 300 résultats (0,47 secondes)

## Rue John Von Neumann, 91400 Orsay

LIMSI, Adresse

Commentaires



Voir les photos

**LIMSI**  
<https://www.limsi.fr/fr/> ▾  
Laboratoire de recherche en informatique pluridisciplinaire, le LIMSI rassemble des chercheurs et enseignants-chercheurs relevant des Sciences de l'Ingénieur ...

**Annuaire - Limsi**  
<https://annuaire.limsi.fr/fr/annuaire> ▾

**LIMSI** ★

Centre National de la Recherche Scientifique

**Adresse :** Rue John Von Neumann, 91400 Orsay  
**Téléphone :** 01 69 85 80 80

Site Web

# Domaines d'application

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

## 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>
```

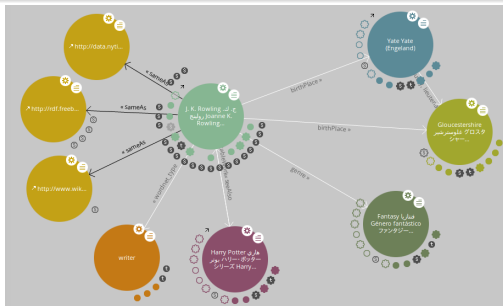
# Quelques exemples de KB

DBPedia

## En résumé

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

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



# Quelques exemples de KB

## Wikidata

### En résumé

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



Main page  
Community portal  
Project chat  
Create a new item  
Item by title  
Recent changes  
Random item  
Query Service  
Nearby  
Help  
Donate

#### Tools

What links here  
Related changes  
Special pages  
Permanent link  
Page information  
Concept URI

Item [Discussion](#)

Read

[View history](#)

## London (Q84)

capital of England and the United Kingdom

[edit](#)

London, UK | London, United Kingdom | London, England

▼ In more languages [Configure](#)

Language	Label	Description	Also known as
English	London	capital of England and the United Kingdom	London, UK London, United Kingdom London, England
French	Londres	capitale du Royaume-Uni	London
Occitan	Londres	No description defined	
Italian	Londra	capitale dell'Inghilterra e del Regno Unito	

[All entered languages](#)

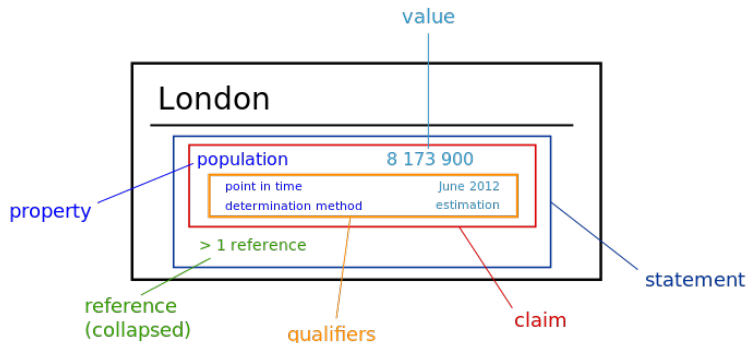
### Statements

# Quelques exemples de KB

Wikidata

## En résumé

- transfert 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 partir 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

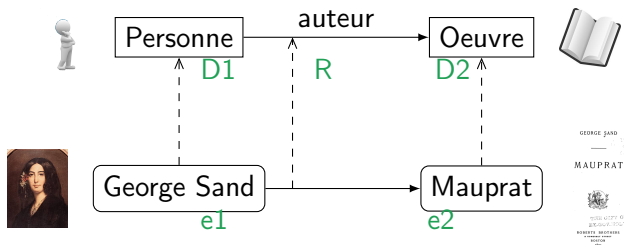
- De quoi parle-t-on ?
  - entités : qui ? quoi ? quand ? où ?
- Qu'en dit-on ?
  - relations entre les entités

# Quelques définitions

relation et types

entités et instance  
de relation

mentions



Mauprat<sub>m1</sub>, que Sand<sub>m2</sub> écrivit<sub>mr</sub> 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<sub>DATE</sub> — elle n'a que 19 ans —, Cecilia Bartoli<sub>PERS</sub> se fait connaître en France.



Cecilia Bartoli



1985

# Étapes

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



Cecilia Bartoli



1985

# Étapes

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



Cecilia Bartoli

se fait connaître



1985

# Étapes

En 1985<sub>DATE</sub> — elle n'a que 19 ans —, Cecilia Bartoli<sub>PERS</sub> 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<sub>DATE</sub> à Washington<sub>LIEU</sub>, George Clooney<sub>PERS</sub> et Barack Obama<sub>PERS</sub> assistent à une conférence de presse sur le Darfour<sub>LIEU</sub>.

# 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

## 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



- désambiguïisation (*entity linking*)

## Exemple de texte

(...) et <personne ref="Barack\_Obama"> Obama </personne> assistent à (...)

# Définition de la tâche : quelles catégories ?

## Quelles catégories ?

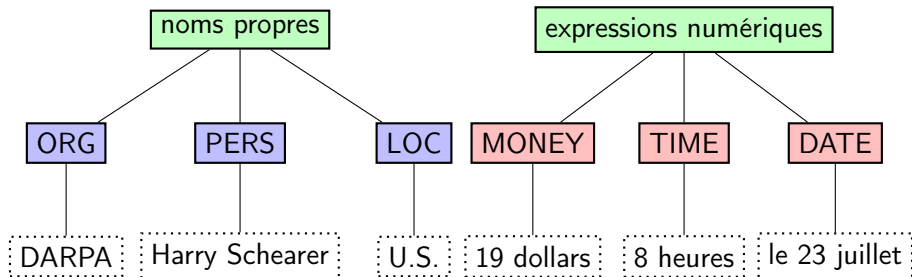
- absence de consensus au-delà des 3 catégories classiques
  - catégorie Divers dans certaines campagnes (CoNLL, HAREM)
- dépendance au type d'application visée
  - granularité des classes : longueur  $\neq$  hauteur
- référence à des jeux existants (campagnes d'évaluation)

## 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<sub>ORG</sub> vote contre un traité d'interdiction des armes nucléaires (ou France<sub>LIEU</sub> ?)

# Catégories d'EN

MUC-6/7





# Catégories d'EN

ACE (2002-2008)

## Caractéristiques

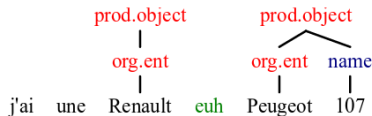
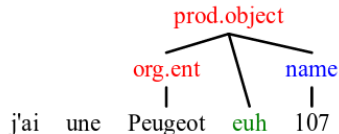
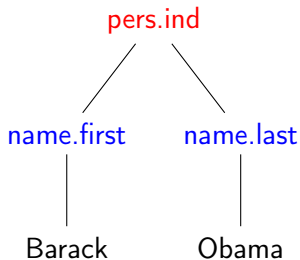
- ajout d'entités non nommées : mentions nominales ou pronominales
- 7 types, dont Person, Organization, Location et :
  - Geo-political Entity
    - **France**<sub>ORG</sub> signed a treaty with Germany last week.
    - The world leaders met in **France**<sub>LIEU</sub> yesterday.
    - **France**<sub>GPE</sub> 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)

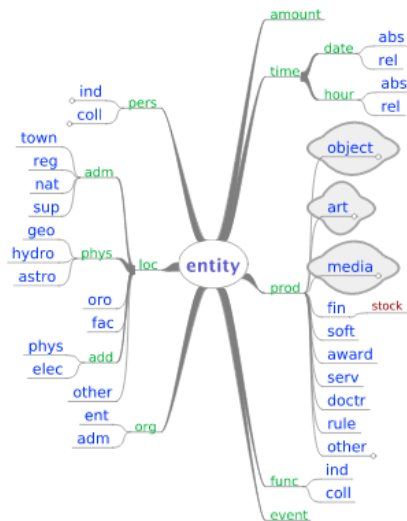
## Caractéristiques

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

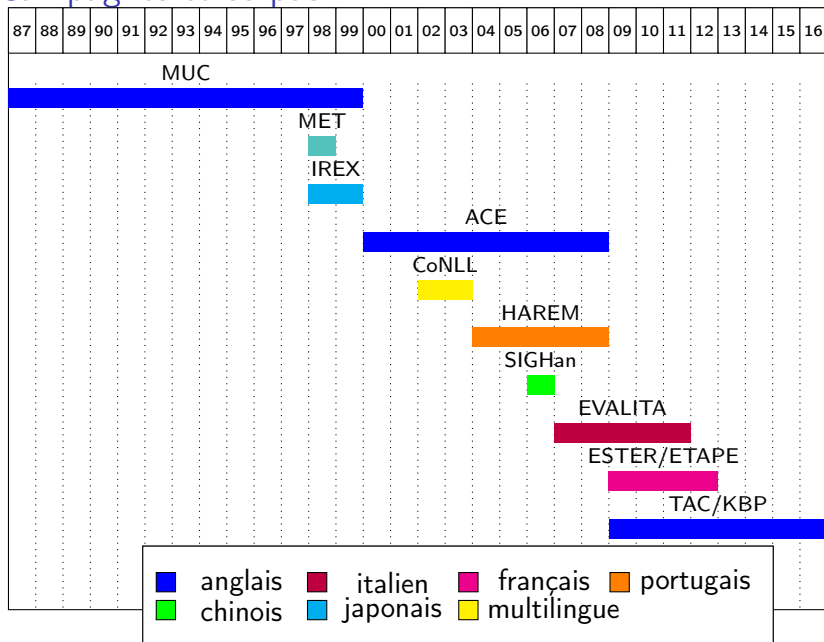


# Catégories d'EN

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<sub>PERS</sub> vs Bill<sub>PERS</sub> and Hillary Clinton<sub>PERS</sub>
- imbrication
  - Université Lyon 2, Comité Exécutif d'Orange
  - Université Lyon 2<sub>ORG</sub> vs Université Lyon<sub>LIEU</sub> 2<sub>ORG</sub> (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  $\Rightarrow$  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)

## 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>.

$$\text{SER} = (0 + 0 + 0 + 0,5 * (1 + 0)) / 2 = 0,25$$

### Hypothèse 2 (système 2)

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

$$\text{SER} = (0 + 0 + 0 + 0,5 * (1 + 1)) / 2 = 0,5$$



# Mon exemple

En 1985<sub>DATE</sub> — elle n'a que 19 ans —, Cecilia Bartoli<sub>PERS</sub> 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**<sub>Hillary\_Clinton</sub> and **Donald J. Trump**<sub>Donald\_Trump</sub> 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**<sub>Hillary\_Clinton</sub>, the **Democratic**<sub>Democratic\_Party\_(United\_States)</sub> candidate, and **Mr. Trump**<sub>Donald\_Trump</sub>, the **Republican**<sub>Republican\_Party\_(United\_States)</sub> 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)

In a grim preview of the discontent that may cloud at least the outset of the next president's term , Hillary Rodham Clinton Hillary Clinton and Donald Trump 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 List of neighborhoods of the District of Columbia by ward 10 eight in 10 voters saying the campaign has left them repulsed rather than excited , the rising toxicity threatens the ultimate victor .

Bill Clinton Democratic Party (United States) Donald Trump Republican Party (United States)  
Mrs. Clinton , the Democratic candidate , and Mr. Trump , the Republican nominee , are seen as dishonest and viewed unfavorably by a majority of voters .

CoreNLP

# Désambiguïsation - définition

## Désambiguïsation/résolution/liaison (*entity linking*)

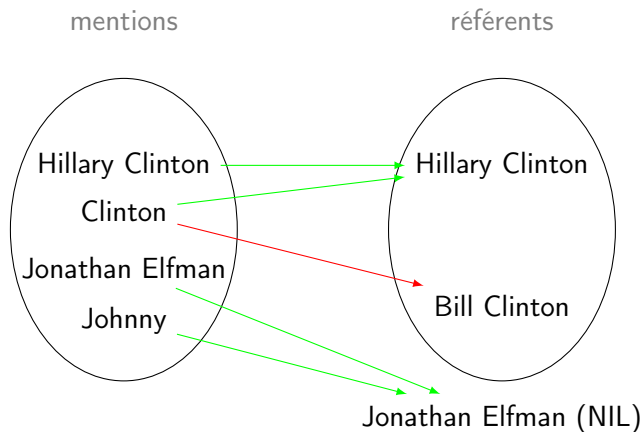
Étant donné une base de connaissances, choisir l'entité correspondant à la mention (référent)

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



## 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

## 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<sub>DATE</sub> — elle n'a que 19 ans —, Cecilia Bartoli<sub>PERS</sub> se fait connaître en France.



Cecilia Bartoli



1985

## 1 Introduction

## 2 Entités

- Définitions
- Reconnaissance d'EN
- Désambiguïsation

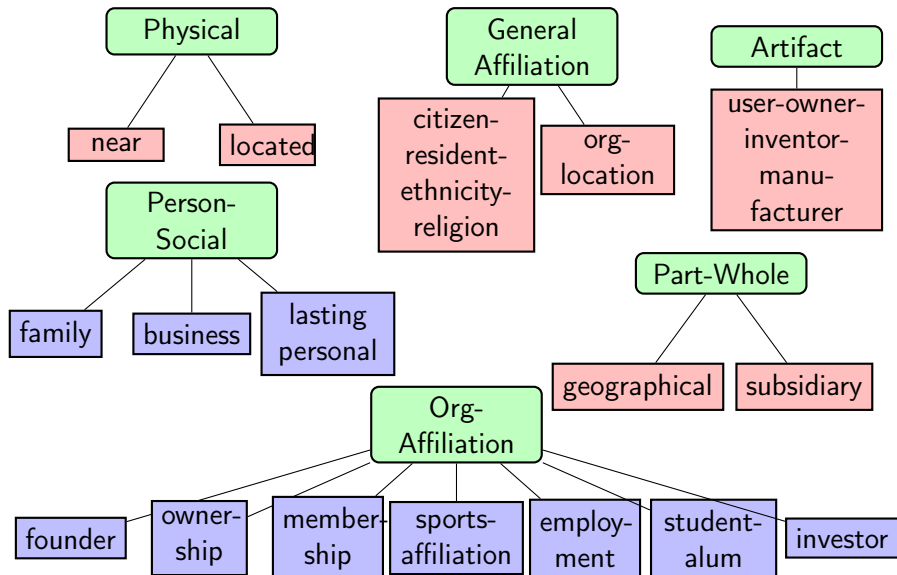
## 3 Relations

- 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

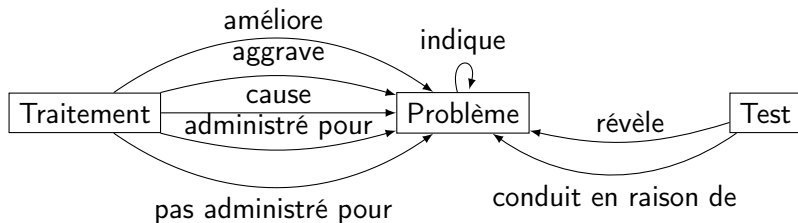
SemEval 2010 tâche 8

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

## 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...

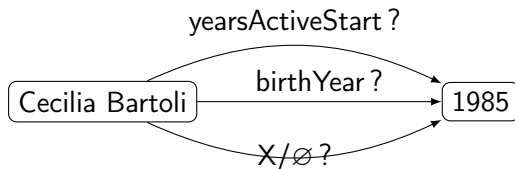
# Extraction de relations

## Définition

Étant donné deux (ou plus) entités, déterminer

- si elles sont en relation
- le type de relation

En 1985<sub>DATE</sub> — elle n'a que 19 ans —, Cecilia Bartoli<sub>PERS</sub> 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

## Cooccurrence

mais ambiguïté

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

## Patrons lexico-syntaxiques

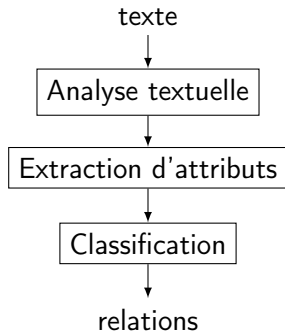
- par exemple pour relation birthyear :
  - *PERSON*, né en *DATE*
  - *PERSON* (*DATE*-)
  - *PERSON* est un GN né en *DATE*
- patrons à écrire manuellement pour chaque relation
  - acquisition automatique possible cependant
  - bootstrapping
- orientés rappel ou précision

# 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

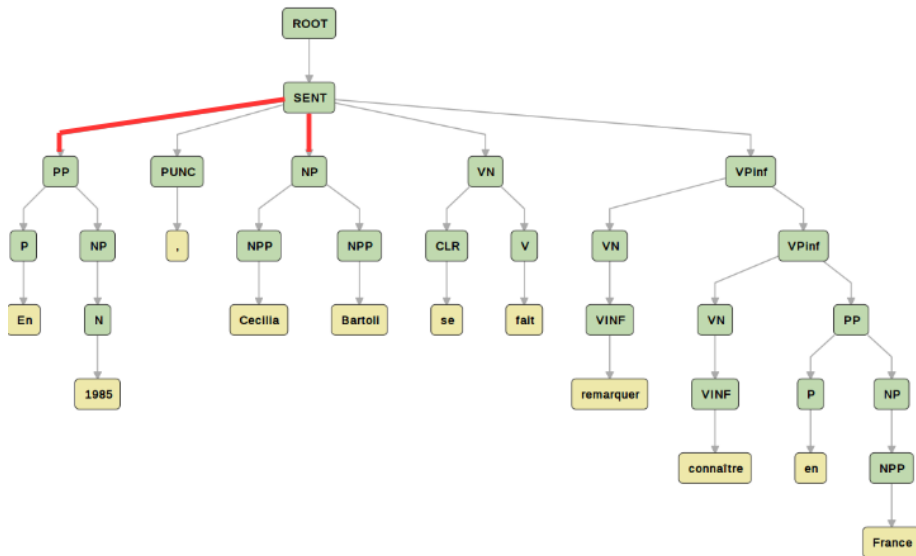
- type  $e_1$  : DATE
- type  $e_2$  : PERSON
- concaténation : DATEPERSON

- syntaxe

- constituants : PP - SENT - NP
- dépendances : nmod -subj

# 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

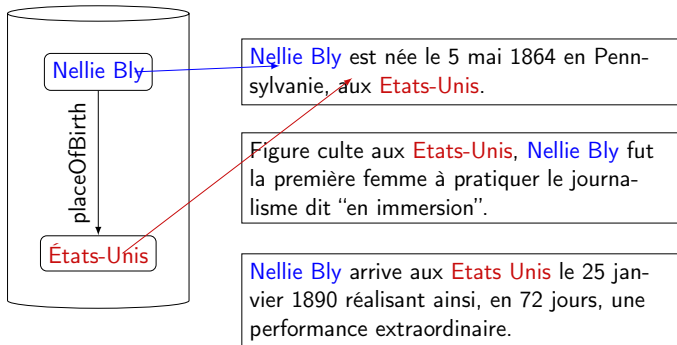
### Utilisation de réseaux de neurones profonds

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

# Supervision distante

## Objectif

- annoter automatiquement les exemples d'entraînement
- ← bases de connaissances
- puis méthodes classiques



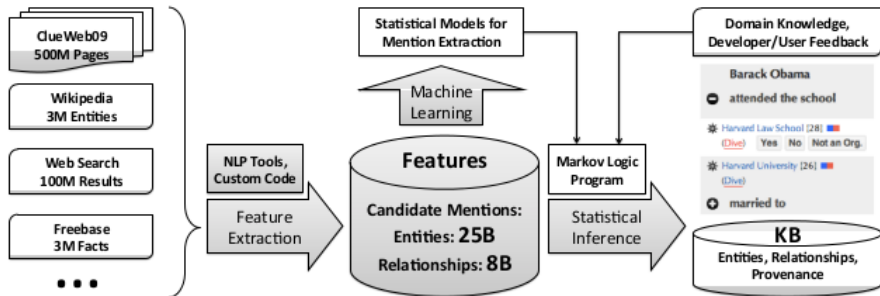
relations dbpedia

## 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.



Ada Lovelace

was one of

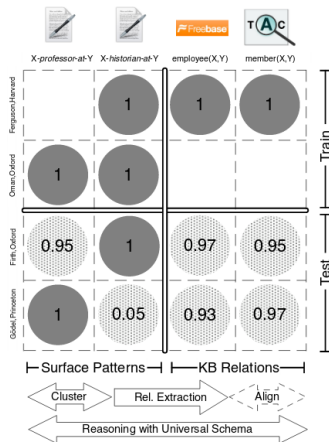
the earliest computer scientist

(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<sub>DATE</sub> — elle n'a que 19 ans —, Cecilia Bartoli<sub>PERS</sub> 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.