Information Extraction

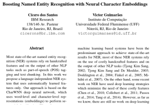
Anne-Laure Ligozat

last update: 2019

NLP & the semantic web

Textual documents and knowledge bases

paper



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

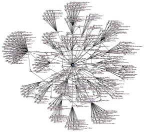
Victor Guimarães
Instituto de Computação
Universidade Federal Fluminense (UFF)
Niterói, RJ, Rio de Janeiro
victorguimaraes@id.uff.br

Abstract

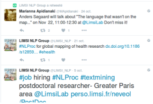
Most state-of-the-art named entity recognition (NER) systems rely on handcrafted features and on the output of other NLP tasks such as part-of-speech (POS) tagging and text chunking. In this work we propose a language-independent NER system that uses automatically learned features only. Our approach 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

wikidata



tweets

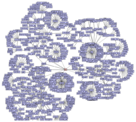


LIMSI NLP Group a retweeté
Marianna Apidianaki @MApidianaki · 24 oct.
Anders Søgaard will talk about "The language that wasn't on the map..." on Nov 22, 11:00-12:30 at @LimsiLab Don't miss it!
2 1

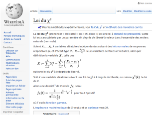
LIMSI NLP Group @LIMSI_NLP · 21 oct.
#NLProc for global mapping of health research dx.doi.org/10.1186/s12859... #ehealth
1

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

wordnet



wikipedia



Wikipédia
L'encyclopédie libre

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 :

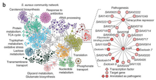
$$f_X(t) = \frac{1}{2^{\frac{k}{2}} \Gamma(\frac{k}{2})} t^{\frac{k}{2}-1} e^{-\frac{t}{2}} \text{ pour tout } t \text{ positif}$$

où Γ est la fonction gamma.

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

Sommaire [masquer]

gene ontology



web page

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

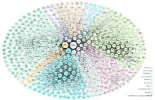
- Violences policières: ce que révèlent les cas Vadot et Jounin
  PAR CHRISTOPHE GUEUGNEAU
- Se disant victime de violences policières, un enseignant de Paris-1 porte plainte
  PAR CHRISTOPHE GUEUGNEAU
- Un sociologue en comparution immédiate
  PAR JOSEPH CONFAVREUX

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.

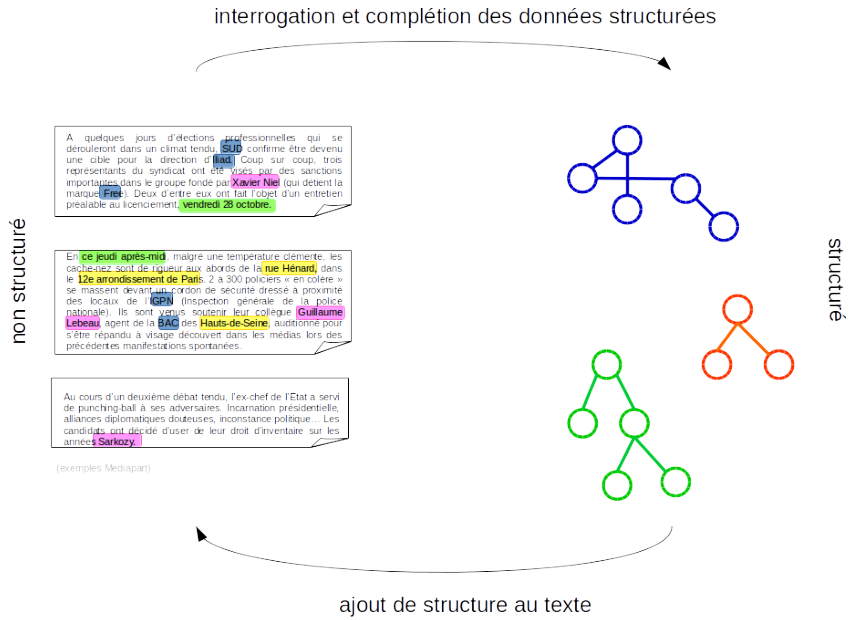
Les faits reprochés à Nicolas Jounin se seraient déroulés Carrefour Pleyel. Selon le récit policier, un groupe aurait chargé un cordon positionné pour les empêcher de rejoindre la Bourse du travail. C'est à cette occasion que Nicolas Jounin aurait foncé sur un policier

…s depuis www.mediapart.fr...

linking open data



Textual documents and structured knowledge



Information/knowledge extraction

Objectives

- targeted understanding of texts

→ produce a structured representation of relevant information

  - relational database
  - knowledge base

→ reasoning and inference

Objective

Machine readable abstract


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

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
...	...	...

Résumé textuel : résumé pour humains

Extraction de connaissances structurées : résumé pour machines

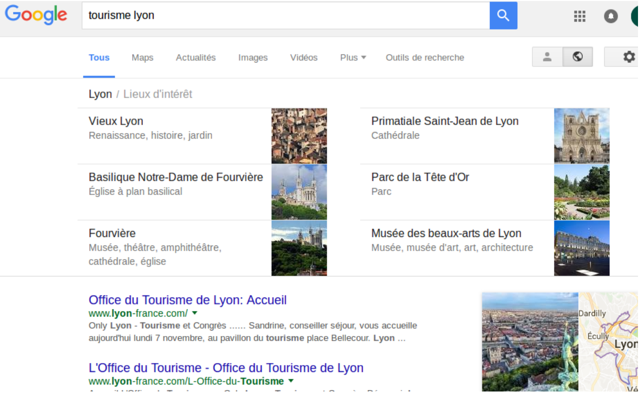
Targeted understanding?

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

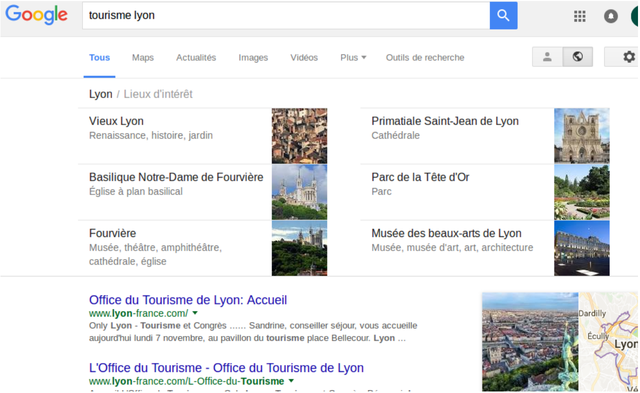
LLNL EQ Lawrence Livermore National Laboratory
LLNL LOC-IN *California*
Livermore LOC-IN *California*
LLNL IS-A scientific research laboratory
LLNL FOUNDED-BY University of California
LLNL FOUNDED-IN 1952

Real life IE applications

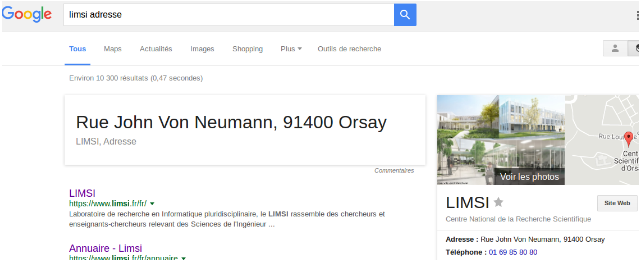
Real life IE applications



Real life IE applications



Real life IE applications



Application domains

- open domain
- digital libraries (google scholar, citeseer)
- bioinformatics
- patent analysis...

Knowledge bases

RDF knowledge base

Set of facts

- fact = sujet, predicate, object
  - resources = entities, concrete or abstract
  - properties = relations, such as height for a Person

Triple example

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

KB examples

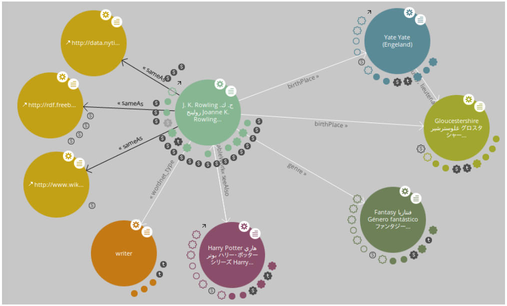
DBPedia

In a nutshell

- more than 4 million RDF resources (nov 2016)
- core = infobox extraction
- ontology partially integrated (created manually)

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

→

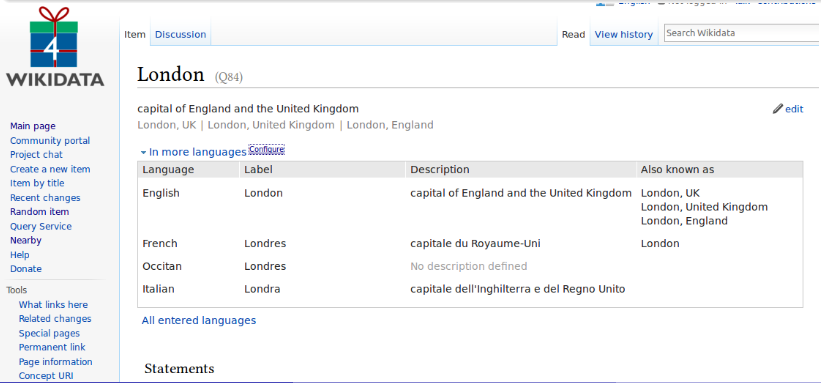


Some KB examples

Wikidata

In a nutshell

- partial tranfer from Freebase
  - collaborative content
  - import of other sources such as MusicBrainz
- more than 20 millions items (nov 2016)

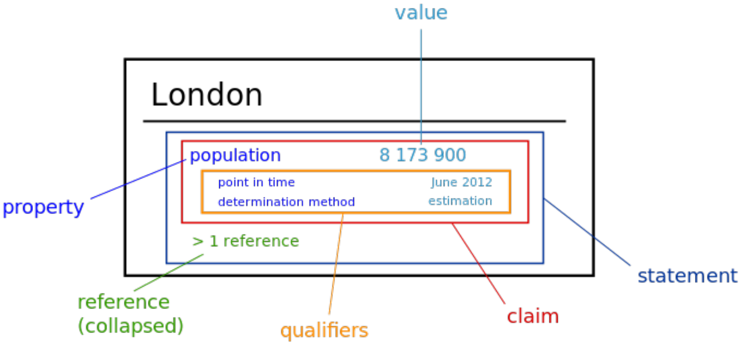


Some KB examples

Wikidata

In a nutshell

- partial tranfer from Freebase
  - collaborative content
  - import of other sources such as MusicBrainz
- more than 20 millions items (nov 2016)



Some KB examples

YAGO

In a nutshell

- ontology with a very rich typing system
- partly built from WordNet and Wikipedia, in particular categories
- built on an RDFS extension
- everything is an entity:
  - objects: cities, persons, URLs, numbers, words...
  - classes (hierarchy)
  - relations
  - facts = (entity, relation, entity)
    - entity $<$ possible to give the reference
- *n*-ary relations = one main fact + other arguments in relation with this fact

Course objective

Information extraction/Knowledge acquisition

- from the NLP view

Main components

- What is it about?

$\rightarrow$ entities: who? what? when? where?

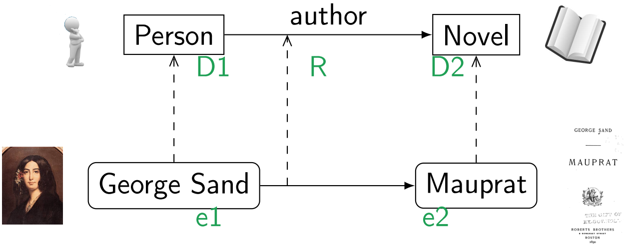
- What is said about it?

$\rightarrow$ relations between entities

A few definitions

relation and types

entities and relation instance



mentions

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

Example
Objective

Find Cecilia Bartoli's first active years in order to store it in a KB (DBPedia relation activeYearsStartYear)

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

Steps

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

Cecilia Bartoli

1985

Steps

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

Cecilia Bartoli

1985

Steps

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

$\boxed{\text{Cecilia Bartoli}} \xrightarrow{\text{se fait connaître}} \boxed{1985}$

Steps

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

Named entity

Definition(s)

- linguistic expression referring to a unique referent in a category in context
- typically: persons, organizations, locations
  - numerical entities often associated: dates, amounts, speed...

Example of an annotated text

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

Annotation formats

- parentheses
  - [$_{ORG}$ 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 or variants (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

NE recognition

Text example

(...) et Obama assistent à (...)

NE recognition

- recognition
  - identification

Text example

(...) et Obama assistent à (...)

NE recognition

- recognition
    - identification

    - categorization

Text example

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

NE recognition

- recognition
  - identification
  - categorization

- entity linking (*désambiguïsation*)

Text example

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

Task definition: which categories?

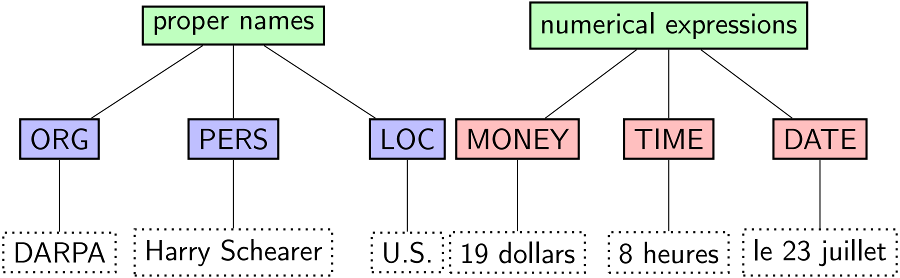
Which categories?

- no consensus beyond the 3 classical categories
  - category Misc in some campaigns (CoNLL, HAREM)
- dependency to the kind of targeted application
  - class granularity: length $\neq$ height
- reference to existing datasets (evaluation campaigns)

Categories ranges?

- which instances?
  - ☺ Matteo Renzi, la famille Kennedy
  - 😑 Zorro, Hercule, les italiens
  - ☹ Mickey, Bison futé, le Prince Charmant
- ambiguity, especially metonymy
  - la France$_{ORG}$ vote contre un traité d'interdiction des armes nucléaires (ou *France*$_{LIEU}$?)

NE categories
MUC-6/7



NE categories

ACE (2002-2008)

Characteristics

- new types of entities: nominal or pronominal mentions
- 7 types, including Person, Organization, Location and:
  - 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)
- hierarchy: subtypes

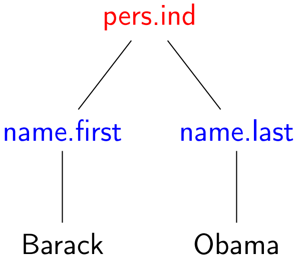
  for example for Person: Individual, Group and Indeterminate (if the context does not enable to disambiguate)

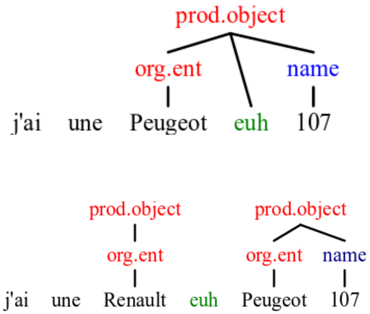
NE categories

Quaero (2011/2012)

Characteristics

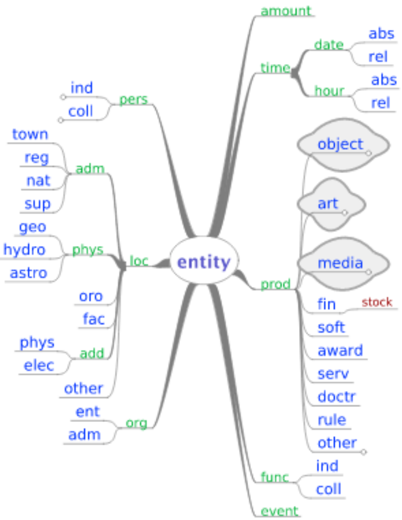
- new types: products, functions
- additional structure: composition
    - metonymy taken into account: two levels of annotation
- annotation adapted to oral corpora (disfluences)



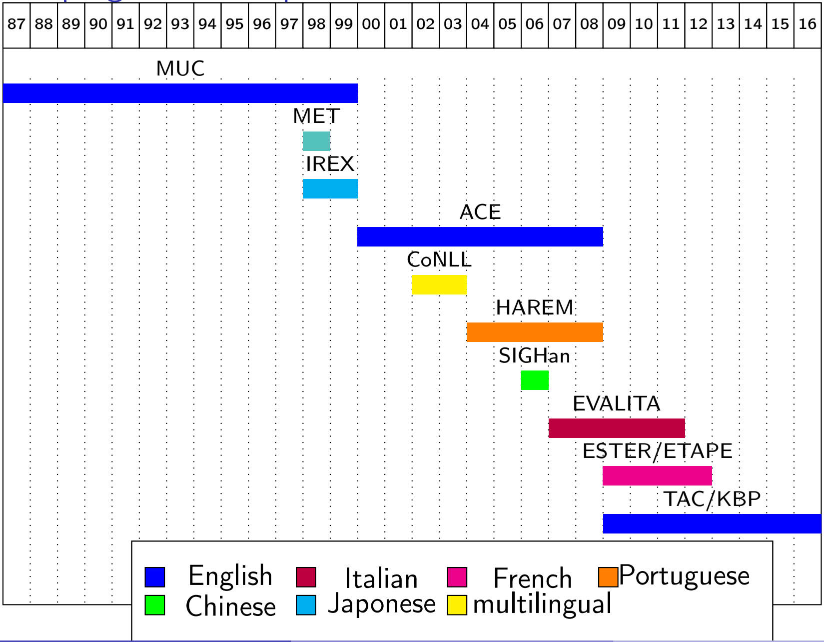


NE categories

Quaero (2011/2012)



Campaigns and corpora



Task definition: which mentions?

Annotation range

- mention forms
  - ☺ proper nouns: *Jacques Chirac*
  - ☹ nicknames, nominal phrases, pronouns: *Chichi*, *l'ancien président*, *il*
- boundaries
  - determiners: *les Rolling Stones*, *La Mecque*, *Le téléphone sonne*
  - functions: *le président Obama*, *l'Abbé Pierre*
  - titles: *Monsieur Fillon*, *Professeur Paolucci*
  - generation: *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$_{ORG}$ (structured entity)

NE recognition

Definition

Automatically identify et classify named entities in texts

Examples of difficultiess

- homonymy (same type or different type)

$\rightarrow$ JFK : person(s) or airport?, Paris

- metonymy

$\rightarrow$ Washington, l'Élysée : location (city) or organization?

Which clues for NE recognition?

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.

Internal features

- case
  - *mRNA* = xXXX, *CPA1* = XXXd
- character *n*-grams
  - *Cotrimoxazole* → drug, *Leuville-sur-Orge* → location
  - *Twilight - Chapitre 3: hésitation* → movie
- words
  - la Banque Populaire
  - l'avenue des Champs-Élysées
  - Benoît XVI (generation item)
- acronym or ampersand
  - Crédit Agricole SA
  - Standard & Poor's
  - F. Hollande
- gazetters (first names for example), word clusters, word embeddings (*plongements lexicaux*)
  - François Hollande

External features

- context of the entity
- additional informations or specific properties
  - Monsieur Hollande
  - Mme Michel
  - Général Leclerc
  - le groupe Sanofi
  - the Coca Cola company
- often given only for the first occurrence of the entity

Symbolic systems

Standard components

- Recognition of triggers and entities from gazetteers
- Cascaded regular expressions

Example of a rule

Université + *de* + CityName ⇒ Organization

Example of an entity recognized by this rule

Université de Nantes

Limits

- low revall: gazetteers incomplete, evolutions, partial entities (*Obama*), noisy texts...
- ambiguities (homonymy and metonymy)

Supervised learning systems

NER as a classification problem

- training
  - create a representative corpus
    - need for many annotated examples!
  - annotate each token
  - choose features adapted to the classes and texts
  - train a classifier to predict the tags from tokens
- test
  - annotate each token
  - evaluate

token	cap	punct	firstname	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

Supervised learning systems

Standard features

- Words
  - current
  - word substrings
  - previous
  - next
- Other learned linguistic information
  - POS tags

Annotation models

- independent tags non adapted
- annotation of tag sequences with a reading sense
  - limits: fixed window, error propagation
- sequence annotation (CRFs)

NER today

Objective

do without a priori knowledge and attribute selection

- deep neural networks [Collobert et al., 2011]

Results

- [Lample et al., 2016]: LSTM-CRF, no external data
- [Guo et al., 2014, Passos et al., 2014]: CRFs + word embeddings
- F1 $\simeq$ 0.90 on CoNLL 2003 data for English (PER, LOC, ORG, MISC)

Evaluation

- tp (true positives) = entities correctly recognized
- fp (false positives) = entities falsely recognized
- fn (false negatives) = entities not recognized

Standard metrics

- Precision $= \frac{tp}{tp+fp}$

$\rightarrow$ entities that were correctly annotated from all entities annotated by the system

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

$\rightarrow$ entities correctly annotated from all entities that should have been annotated

Evaluation

Reference

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

Hypothesis (system output)

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

Disadvantage of these metrics for named entities

- *Jean-Yves* falsely recognized as an entity

→ false positive

- *Jean-Yves Le Drian* not recognized

→ false negative

Adapted metrics

- R : # entities in the reference
- H : # entities in hypothesis (= system output)
- C : # correct entities (= true positives)
- T : # entities with correct boundaries but wrong category
- F : # entities with correct category but wrong boundaries
- TF : # entities with wrong type and boundaries
- I : # inserted entities insérées (= false positives)
- D : # forgotten entities (= false negatives)

Adapted metrics

→ partial recognition = half correct

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

Evaluation example

Reference (manual annotation)

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

Hypothesis 1 (system 1)

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

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

Hypothesis 2 (system 2)

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

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

My example

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

Cecilia Bartoli

1985

Entity linking - definition

Entity linking (*désambiguïsation/résolution/liaison*)

Given a knowledge base, chose the entity corresponding to the mention (*referent*)

Text to analyze

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.

Entity linking - definition

Entity linking (*désambiguïsation/résolution/liaison*)

Given a knowledge base, chose the entity corresponding to the mention (*referent*)

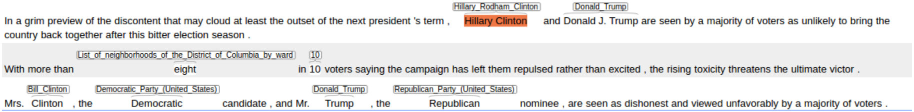
Expected result

In a grim preview of the discontent that may cloud at least the outset of the next president's term, Hillary Clinton$_{\textit{Hillary_Clinton}}$ and Donald J. Trump$_{\textit{Donald_Trump}}$ 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$_{\textit{Hillary_Clinton}}$, the Democratic$_{\textit{Democratic_Party_(United_States)}}$ candidate, and Mr. Trump$_{\textit{Donald_Trump}}$, the Republican$_{\textit{Republican_Party_(United_States)}}$ nominee, are seen as dishonest and viewed unfavorably by a majority of voters.

Entity linking - definition

Entity linking (*désambiguïsation/résolution/liaison*)

Given a knowledge base, chose the entity corresponding to the mention (*referent*)



CoreNLP

Entity linking - definition

Entity linking (*désambiguïsation/résolution/liaison*)

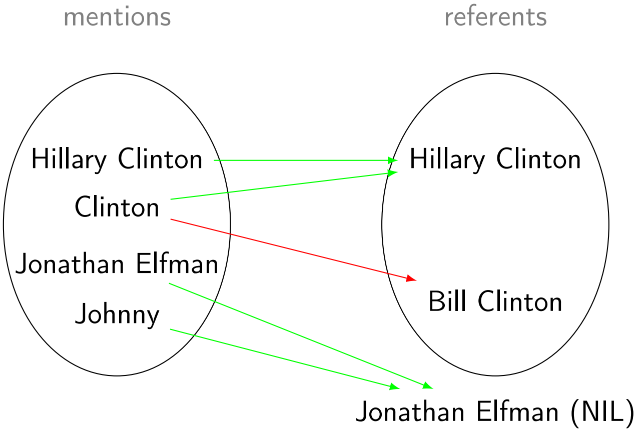
Given a knowledge base, chose the entity corresponding to the mention (*referent*)

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

Entity linking - difficulties



Entity linking - principles

Steps

1. possible mention detection
    - often based on NE recognition
2. candidate selection
    - graphical proximity to labels, links texts, querys that lead to the Wikipedia pages, Wikipedia disambiguation pages
3. candidate ranking

   WSD / Wikipedia

    - mention: distance to the referents' labels
    - referent: popularity (most frequent, Wikipedia page with most links...)
    - local context of the mention: textual similarity with Wikipedia pages, links...
    - global context of the mention (document): other entities (collective disambiguation), coreference

KBP

Tâche Entity Discovery and Linking

- Discovery: detect and annotate mentions
  - classes: LOC, ORG, PER, FAC, GPE;
  - mentions: EN, noms, posts authors
- Linking: attach mention clusters to a KB
- Difficulties (KBP 2015):
  - detection of common names and acronyms
  - rare entities
  - popularity bias
  - general knowledge
  - informal language
  - lack of coherence between NE type and referent
- F1 $\simeq$ 0.60 for EL in English 2015

My example

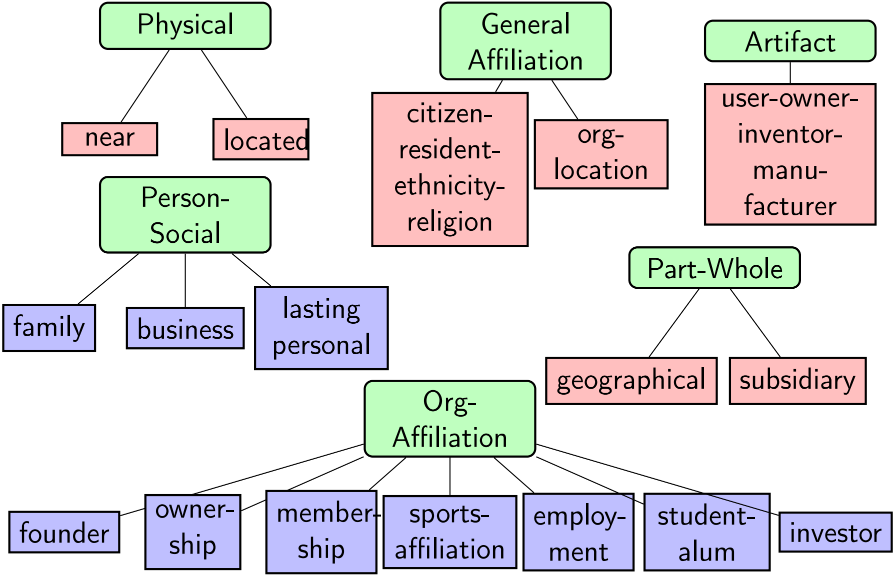
En 1985$_{DATE}$ — elle n'a que 19 ans —, Cecilia Bartoli$_{PERS}$ se fait connaître en France.

Cecilia Bartoli

1985

A few sets of relations

ACE 2005



A few sets of relations

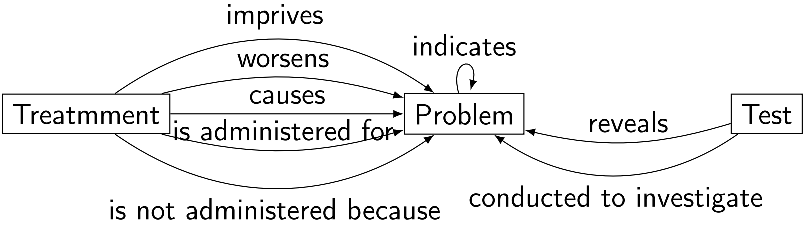
SemEval 2010 task 8

Type	Example
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.

A few sets of relations

i2b2 2010

Natural Language Processing Challenge for Clinical Records



A few sets of relations

Freebase

Most frequent Freebase relations

- /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

Relation characteristics

- between concepts or concept instances
- hierarchical or not
- include events or only binary relations
- “real world” relations or factivity taken into account

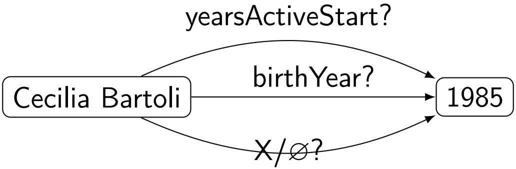
Relation extraction

Definition

Given two (or more) entities, determine

- whether there is a relation between them
- what kind of relation

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



Difficulties

Variabiliy of expression of 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.

wikipédia, les échos et encyclopédie universalis examples

Simple methods

Cooccurrence

but ambiguity

- person - date: start date of the career, birthdate, other?
- treatment - illness: cures? prevents? side effect?

Lexico-syntactic patterns

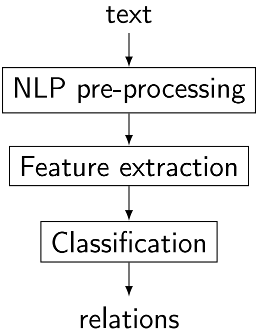
- for example for birthyear relation:
  - *PERSON*, born in *DATE*
  - *PERSON* (*DATE*-)
  - *PERSON* is a NP born in *DATE*
- have to be manually written for each relation
  - possibile automatic acquisition
  - bootstrapping
- oriented recall or precision

Supervised relation extraction

Classification problem

- binary or multiclass classification
- positive and negative training examples

Basic supervised method



Context

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

Features

- words (or lemmss)
  - from different context parts
  - bag of word and n-grams
  - syntactic head and concatenation
- entity types
  - entity types and concatenation
- syntactic information
  - constituents path
  - dependency path
- external resources
  - country list, triggers...

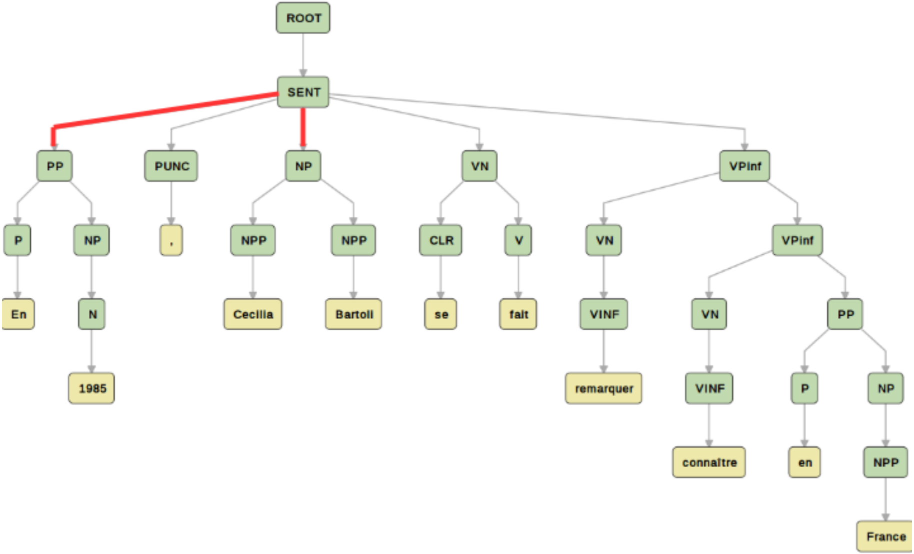
Feature example

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

- words
  - before e_1: {En}
  - between entities (bow): {elle, n', a, que 19, ans}
  - after e_2: {se, fait, connaître, en, France}
  - head e_1: 1985
  - head e_2: Bartoli
- types
  - type e_1: DATE
  - type e_2: PERSON
  - concatenation: DATEPERSON
- syntax
  - constituents: PP - SENT - NP
  - dependencies: nmod -suj

Example

Constituents parse tree



Structured representations

Which attributes?

- intuition
- experiments

Using structured representations

Definition of adapted similarity metrics: parse tree kernels

Experiments

- constituent tree [Zelenko et al., 2003]
- dependency tree [Culotta and Sorensen, 2004]
- shortest path between entities [Bunescu and Mooney, 2005]

Limits of these approaches

Disadvantages of previous methods

- classify quality strongly dependent on pre-processing
- large annotated corpora
  - event if crowdsourcing possible [Liu et al., 2016]
- corpora imbalance
- lack of generalization

Getting rid of pre-processing

Deep neural networks

- input: words, *n*-grams + positions + word embeddings
- network: RNN or CNN

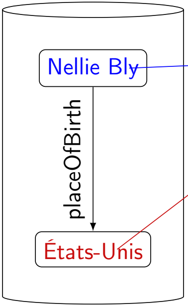
Distant supervision

Objective

- automatically annotate training examples

← KBs

- then standard methods



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

Distant supervision

Hypotheses

- any sentence that contains a pair of entities that participate in a known KB relation is likely to express that relation in some way [Mintz et al., 2009, Wu and Weld, 2007, Niu et al., 2012]
- multi-instances learning problem [Riedel et al., 2010]: at leat one of the sentences contains a relation mention
- Several relations may exist between two entities [Hoffmann et al., 2011]

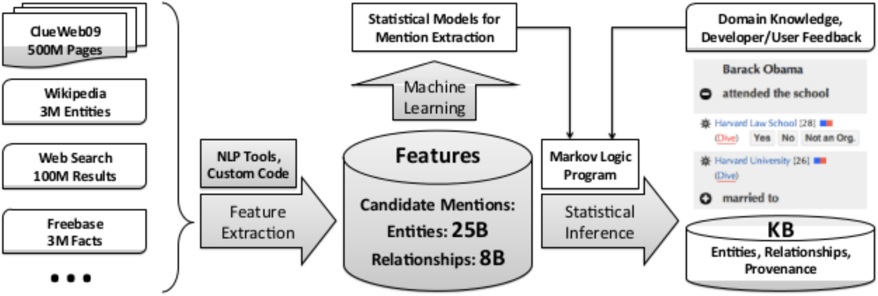
☺ (mostly) domain independent

☺ scales up well

☹ only for context free relations

☹ depends on the quality of NER

DeepDive [Niu et al., 2012]



Open information extraction

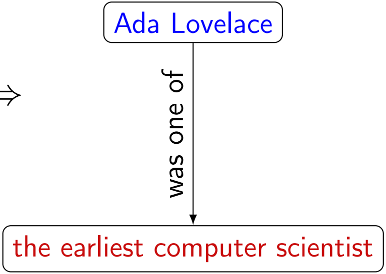
Principle

Keep the relation expression from the text

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.

$\Rightarrow$

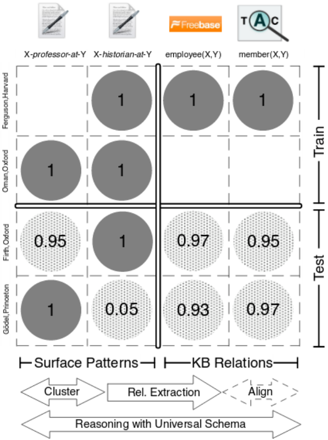


(exemples de `http://openie.allenai.org/`)

Open information extraction

Limits

- Non normalized relations
  - [Angeli et al., 2015]: cooccurrences relations OIE and KBP in corpus
  - [Riedel et al., 2013, Verga et al., 2016]: implications between relations



Current difficulties

- rare relations (in texts)
- contextual relations
- factivity
- source: reliability, fiction...
- common sense knowledge
- NL qury: several relations

My example

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

Cecilia Bartoli $\xrightarrow{\text{yearsActiveStart}}$ 1985

☺

Conclusion

A few points

- more and more explicit knowledge
- virtuous circle between IE and semantic annotation
- remain complementary
  - query both types of resources
  - real interaction between reasoning on texts and knowledge

References

- 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.