

# ***I.A. et Langage :*** **Traitement automatique du** **langage naturel**

**Elena CABRIO**

[elena.cabrio@univ-cotedazur.fr](mailto:elena.cabrio@univ-cotedazur.fr)

**Serena VILLATA**

[villata@i3s.unice.fr](mailto:villata@i3s.unice.fr)

# **Extraction d'information et Reconnaissance d'Entités nommées**

# Information non structurée vs structurée

La plupart des données (aussi sur le Web) ne sont pas structurées:

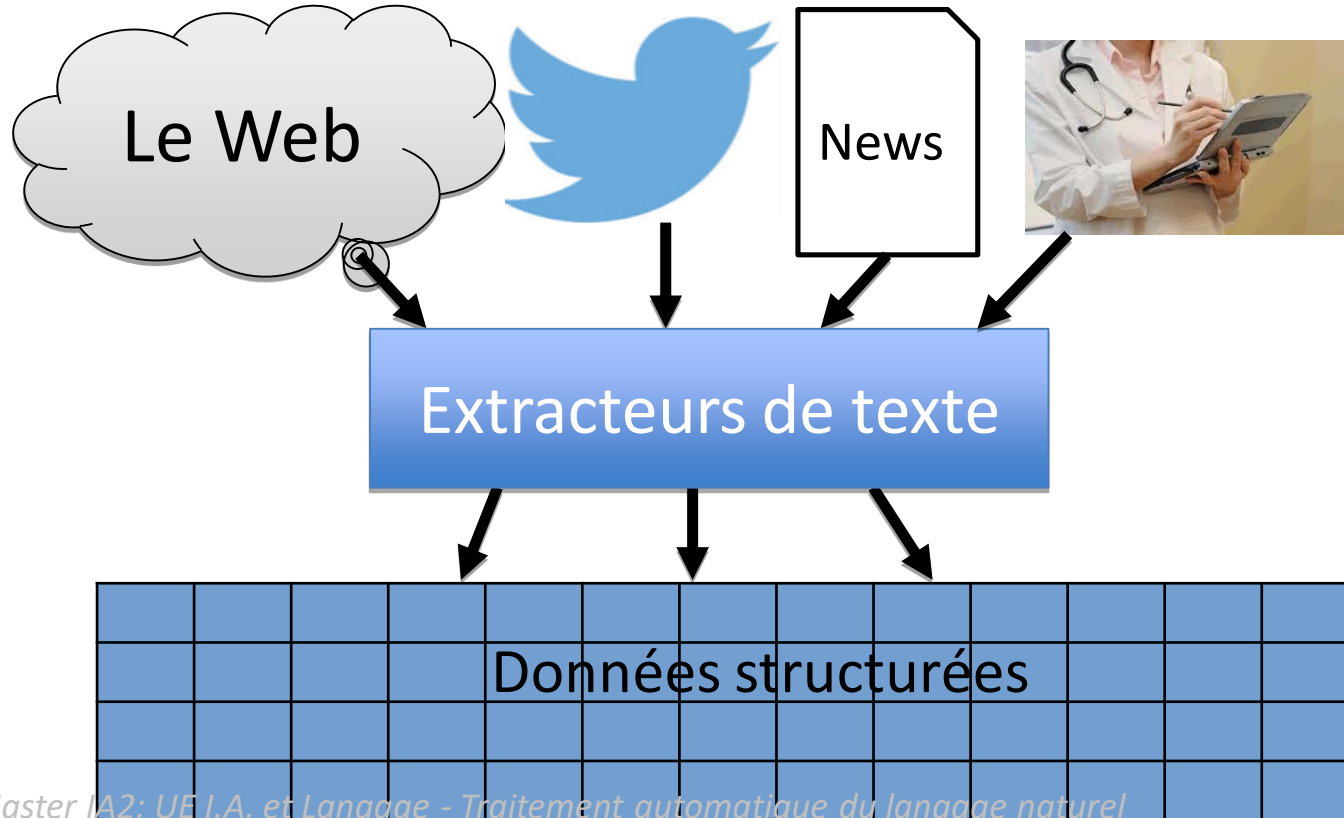
- Texte
- Speech
- Images



facebook



# Extraire la connaissance d'un texte



# Information extraction

---

From Wikipedia, the free encyclopedia

**Information extraction** (IE) is the task of automatically extracting structured information from **unstructured** and/or semi-structured **machine-readable** documents. In most of the cases this activity concerns processing human language texts by means of **natural language processing** (NLP). Recent activities in multimedia document processing like automatic annotation and content extraction out of images/audio/video could be seen as information extraction.

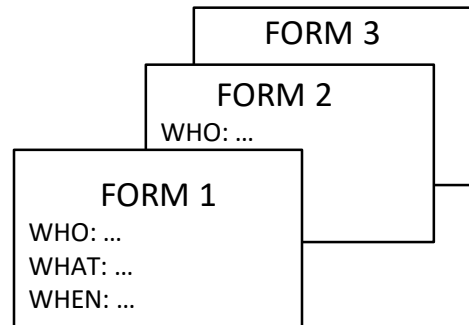
# Extraction d'information

- Systèmes d'extraction d'informations (IE):
  - Trouver et comprendre des parties limitées et pertinentes de texte
  - Recueillir des informations à partir de nombreux morceaux de texte
  - Produire une représentation structurée des informations pertinentes:
    - relations (au sens de la base de données)
    - une base de connaissances
- **Buts:**
  - Organiser l'information pour qu'elle soit utile aux gens
  - Mettre les informations sous une forme sémantiquement précise qui permet de faire d'autres déductions à l'aide d'algorithmes informatiques

# Extraction d'information

Les systèmes IE extraient de l' **information claire, factuelle**

*Qui a fait quoi à qui quand?*

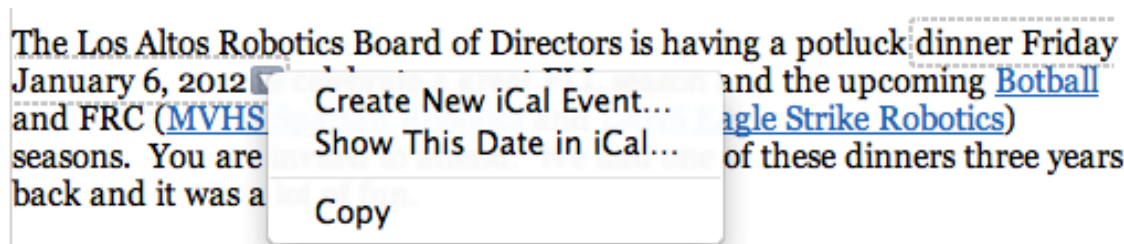


Mappage de textes dans un format fixe  
(modèles) représentant les informations clés

# Example (simple):

## Extraction d'information de bas niveau

Disponible dans des applications telles que Apple ou Google Mail et l'indexation Web



Souvent basé sur des expressions régulières et des listes de noms



# An example (simple)



Vente Villa 4 pièces Nice (06000)  
Réf. 12390: Sur les Hauteurs de Nice. Superbe villa moderne (190m<sup>2</sup>), 2 chambres et 1 suite parentale, 3 salles de bain. Très grand salon/salle à manger, cuisine américaine équipée. Prestations de haut standing. Vue panoramique sur la mer. Cette villa a été construite en 2005. 1 270 000 euros. Si vous êtes intéressés, contactez vite Mimi LASOURIS. 06.43.43.43. 43



## Agence immobilière

Reference: ???

Prix: ???

Surface: ???

Année de construction: ???

Pièces: ???

Propriétaire: ???

Téléphone: ???

# An example (simple)



Vente Villa 4 pièces Nice (06000)  
Réf. 12390: Sur les Hauteurs de Nice. Superbe  
villa moderne (190m<sup>2</sup>), 2 chambres et 1 suite  
parentale, 3 salles de bain. Très grand salon/salle  
à manger, cuisine américaine équipée.  
Prestations de haut standing. Vue panoramique  
sur la mer. Cette villa a été construite en 2005 1  
270 000 euros. Si vous êtes intéressés, contactez  
vite Mimi LASOURIS. 06.43.43.43. 43

## Agence immobilière

Reference: ???

Prix: ???

Surface: ???

Année de construction: ???

Pièces: ???

Propriétaire: ???

Téléphone: ???

# An example plus complex...



*“Yess! Yess! Its official Nintendo announced today that  
they Will release the Nintendo 3DS in north America  
march 27 for \$250”*

# An example plus complex...



*“Yess! Yess! Its official **Nintendo** announced today that they Will release the **Nintendo 3DS** in **north America** **march 27** for **\$250**”*

# An example plus complex...



*“Yess! Yess! Its official **Nintendo** announced today that they Will release the **Nintendo 3DS** in **north America** **march 27** for **\$250**”*

COMPANY	PRODUCT	DATE	PRICE	REGION

**PRODUCT RELEASE**

# An example plus complex...



*"Yess! Yess! Its official **Nintendo** announced today that they Will release the **Nintendo 3DS** in **north America** **march 27** for **\$250**"*

COMPANY	PRODUCT	DATE	PRICE	REGION
Nintendo	3DS	March 27	\$250	North America

## PRODUCT RELEASE

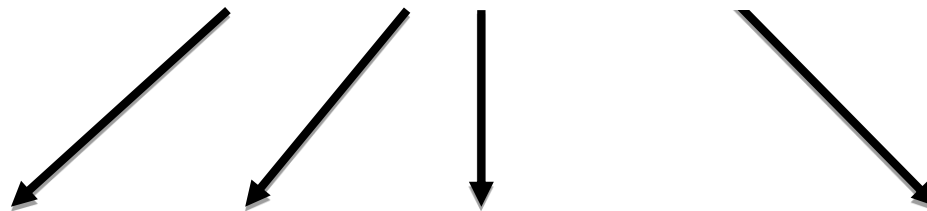
# An example plus complex...



***Samsung Galaxy S5 Coming to All Major U.S. Carriers  
Beginning April 11th***

COMPANY	PRODUCT	DATE	PRICE	REGION
Samsung	Galaxy S5	April 11	?	U.S.
Nintendo	3DS	March 27	\$250	North America

# An example plus complex...



COMPANY	PRODUCT	DATE	PRICE	REGION
Samsung	Galaxy S5	April 11	?	U.S.
Nintendo	3DS	March 27	\$250	North America
...	...	...	...	...

## PRODUCT RELEASE



# Applications possibles

- **Question Réponse**

- *Quelles entreprises lancent de nouveaux smartphones en Europe ce printemps?*
- *Alertez-moi à chaque fois qu'un nouveau smartphone est annoncé aux États-Unis.*

- **Fouille des données**

- *Analyser les tendances des versions de produits dans différents secteurs*
- *Existe-t-il une corrélation entre le prix et la date de sortie?*

# IE vs Recherche d'information

## Recherche d'information (IR)

- *Requete de l'utilisateur* → Textes pertinents
- *Approche*: correspondance des mots clés
- *Généralité de la requête* : totale



## Extraction d'information (IE)

- *Analyse linguistique ciblée sur des informations pertinentes*
- *Requete de l'utilisateur* → Information pertinente
- *Approche*: Analyse linguistique
- *Généralité de la requête* : limité à l'information cible

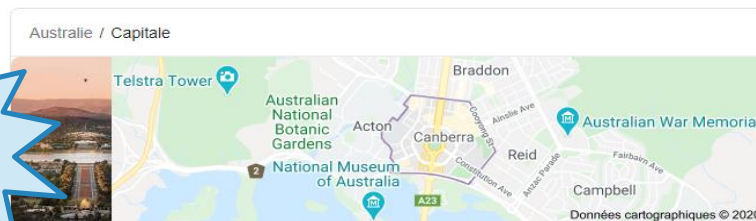
# IE vs Recherche d'information



quelle est la capitale de l'australie

Tous Maps Actualités Images Shopping Plus Paramètres Outils

Environ 9 030 000 résultats (0,98 secondes)



Canberra

Recherches associées

Voir d'autres éléments (plus de 10)



Australie



Territoire de  
la ca...



Sydney



Melbourne



Perth



Ottawa



Brisbane

Signaler un problème

fr.wikipedia.org › wiki › Canberra

Canberra — Wikipédia

Canberra /kænbə/ (en anglais : /kænbe/) est la **capitale de l'Australie** et du Territoire de la **capitale australienne**. La ville est située à l'extrémité sud du ...

Population: 426 704 hab. (2019)

Maire: Jon Stanhope

État: Territoire de la capitale australienne

Superficie: 814,2 km<sup>2</sup>

Géographie · Histoire · Gouvernement · Infrastructures

fr.wikipedia.org › wiki › Australie

Australie — Wikipédia

Sa capitale est Canberra, située dans le Territoire de la **capitale australienne**. Sa population estimée à 25,6 millions d'habitants en mars 2020, est principalement ...

Capitale: Canberra 35° 15' S, 149° 08' E

Langues officielles: Anglais (de facto)

Forme de l'État: Monarchie constitutionnelle ...

Gouverneur général: David Hurley

Étymologie · Géographie · Histoire

voyageuraustralien.fr › capitale-australie

Quelle est la capitale de l'Australie? - Guide du voyageur ...

L'**Australie** est une souveraineté, à la fois continent et pays, localisée dans la partie sud-est du globe. Elle comprend le territoire principal **Australien**, l'île de ...

www.laculturegenerale.com › capitale-australie

Quelle est la capitale de l'Australie ? | La culture générale

25 juin 2018 - **capitale australie** canberra, **capitale australie** sydney, brisbane, géographique, question, la plus grande ville d'Australie.

IE

IR

# IE vs Recherche d'information



Google

quelle est la capitale de l'australie

Tous Maps Actualités Images Shopping Plus Paramètres Outils

Environ 9 030 000 résultats (0,98 secondes)

Australie / Capitale

Canberra

Recherches associées

Voir d'autres éléments (plus de 10)

Australie Territoire de la ca... Sydney Melbourne Perth Ottawa Brisbane

Signaler un problème

fr.wikipedia.org › wiki › Canberra ▼

## Canberra — Wikipédia

Canberra /kænbə/ (en anglais : /kænbe/) **est la capitale de l'Australie** et du Territoire de la **capitale australienne**. La ville **est** située à l'extrémité sud du ...

Population: 426 704 hab. (2019)

Maire: Jon Stanhope

État: Territoire de la capitale australienne

Superficie: 814,2 km<sup>2</sup>

Géographie · Histoire · Gouvernement · Infrastructures

fr.wikipedia.org › wiki › Australie ▼

## Australie — Wikipédia

Sa capitale **est** Canberra, située dans le Territoire de la **capitale australienne**. Sa population estimée à 25,6 millions d'habitants en mars 2020, **est** principalement ...

Capitale: Canberra 35° 15' S, 149° 08' E

Langues officielles: Anglais (de facto)

Forme de l'État: Monarchie constitutionnelle ...

Gouverneur général: David Hurley

Étymologie · Géographie · Histoire

voyageuraustralien.fr › capitale-australie ▼

## Quelle est la capitale de l'Australie? - Guide du voyageur ...

L'**Australie** **est** une souveraineté, à la fois continent et pays, localisée dans la partie sud-est du globe. Elle comprend le territoire principal **Australien**, l'île de ...

www.laculturegenerale.com › capitale-australie ▼

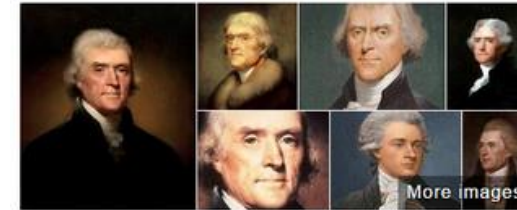
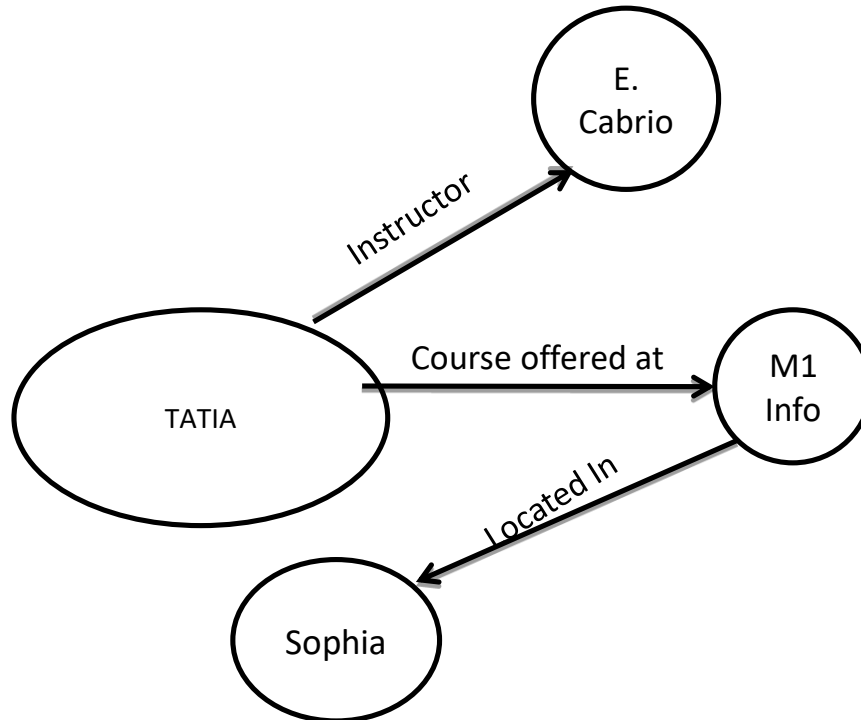
## Quelle est la capitale de l'Australie ? | La culture générale

25 juin 2018 - **capitale australie** canberra, **capitale australie** sydney, brisbane, géographique, question, la plus grande ville d'Australie.



# Graphes de connaissance

Des “objets” pas de chaines!



## Thomas Jefferson

3rd U.S. President

Thomas Jefferson was an American Founding Father; the principal author of the Declaration of Independence, and the third President of the United States. [Wikipedia](#)

**Born:** April 13, 1743, Shadwell, VA

**Died:** July 4, 1826, Charlottesville, VA

**Presidential term:** March 4, 1801 – March 4, 1809

**Spouse:** [Martha Jefferson](#) (m. 1772–1782)

**Party:** [Democratic-Republican Party](#)

**Awards:** [AIA Gold Medal](#)

Get updates about Thomas Jefferson

People also search for

[View 15+ more](#)



John Adams



George Washington



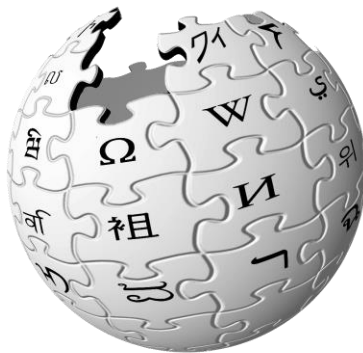
Benjamin Franklin



James Madison



Alexander Hamilton



WIKIPEDIA  
*The Free Encyclopedia*

# Sources de données

Crostata



Apple crostata with slivered almonds

Origin

```
{{Infobox prepared food
| name                = Crostata
| image               = [[File:Crostata di mele e mandorle di Adriano a profilo.jpg|300px]]
| imagesize          = 
| caption             = Apple crostata with slivered almonds
| alternate_name      = 
| country             = [[Italy]]
| region              = 
| creator             = 
| course              = [[Dessert]]
| type                = [[Tart]]
| served              = 
| main_ingredient    = Pastry crust, [[jam]] or [[ricotta]], fruit
| variations          = ''Crostata di frutta'', ''crostata di ricotta'', many other sweet or savoury variations
| calories            = 
| other               = 
}}
```



WIKIPEDIA  
*The Free Encyclopedia*

# Sources de données

```
{{Infobox prepared food
| name                = Crostata
| image               = [[File:Crostata di mele e mandorle di Adriano
| imagesize           =
| caption             = Apple crostata with slivered almonds
| alternate_name      =
| country             = [[Italy]]
| region              =
| creator              =
| course              = [[Dessert]]
| type                = [[Tart]]
| served              =
| main_ingredient     = Pastry crust, [[jam]] or [[ricotta]], fruit
| variations          = ''Crostata di frutta'', ''crostata di ricotta'', many other sweet or savoury variations
| calories             =
| other               =
}}
```

Dessert



A flourless chocolate cake (torte)

**Type** Usually sweet

**Variations** Numerous (biscuits, cakes, tarts, cookies, gelatins, ice creams, pastries, pies, puddings, custards, and sweet soups, etc.)

Cookbook: Dessert Media: Dessert






WIKIPEDIA  
*The Free Encyclopedia*

# Sources de données

```
{{Infobox prepared food
| name                = Crostata
| image               = [[File:Crostata di mele e mandorle di Adriano
| imagesize           =
| caption             = Apple crostata with slivered almonds
| alternate_name      =
| country             = [[Italy]]
| region              =
| creator              =
| course              = [[Dessert]]
| type                = [[Tart]]
| served              =
| main_ingredient     = Pastry crust, [[jam]] or [[ricotta]], fruit
| variations          = ''Crostata di frutta'', ''crostata di ricotta'', many other sweet or savoury variations
| calories            =
| other               =
}}
```

**Dessert**

**Tart**



**Type** [Blueberry tart](#)

**Variations**

**Main ingredients** [Pastry crust \(usually shortcrust pastry\)](#)

**Variations** [Sweet tarts, savoury tarts](#)

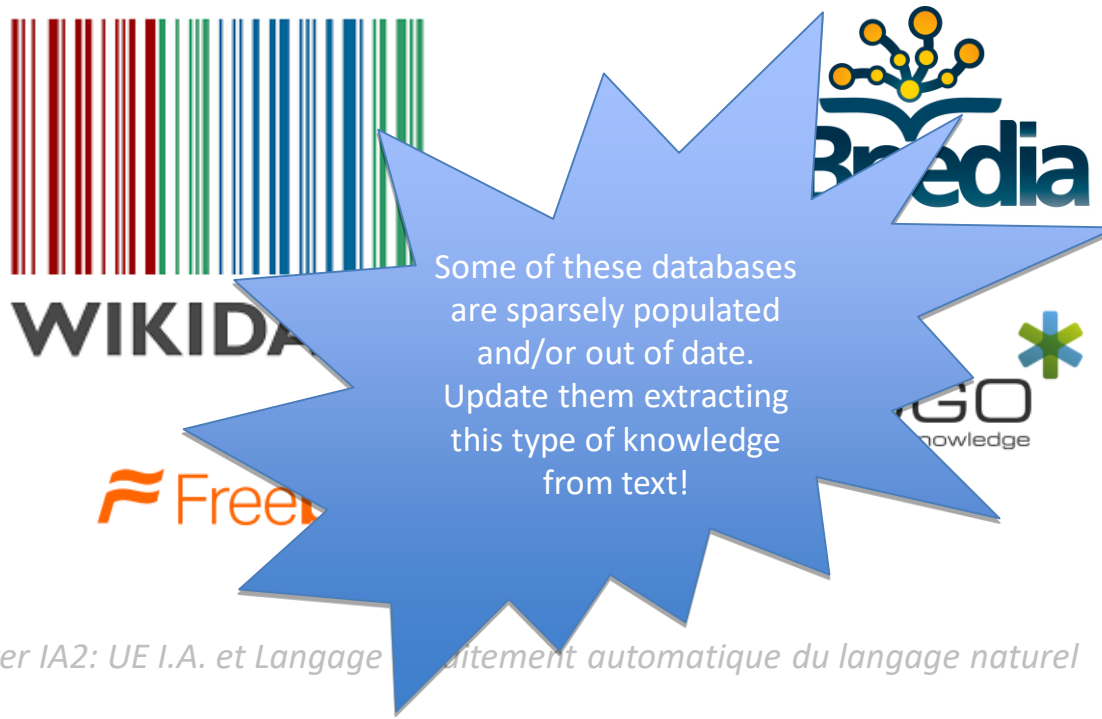
[Cookbook: Tart](#) [Media: Tart](#)



# Sources de données disponibles



# Sources de données disponibles



# Peuplement de la base de connaissances: sous tâches

- Reconnaissance d'entités / Classification / Linking
- Extraction des relations
- Extraction d'événements
- Inférence sur la base de connaissances

# Reconnaissance d'entités nommées

La **reconnaissance d'entités nommées** est une sous-tâche de l'activité d'extraction d'information dans des corpus documentaires. Elle consiste à rechercher des objets textuels (c'est-à-dire un mot, ou un groupe de mots) catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates, etc.

## Identification et catégorisation

- **Entités** (personnes, organisations, lieu)
- **Temps** (dates, temps et durée)
- **Quantité** (valeurs monétaires, mesures, pourcentages et nombres cardinaux)

*La décision du député indépendant Andrew Wilkie de retirer son soutien au gouvernement Travailliste minoritaire a semblé dramatique, mais cela ne devrait pas menacer davantage sa stabilité. Lorsque, après l'élection de 2010, Wilkie, Rob Oakeshott, Tony Windsor et les Verts ont décidé de soutenir le parti Travailliste, ils n'ont donné que deux garanties: la confiance et l'approvisionnement.*

# Reconnaissance d'entités nommées

La **reconnaissance d'entités nommées** est une sous-tâche de l'activité d'extraction d'information dans des corpus documentaires. Elle consiste à rechercher des objets textuels (c'est-à-dire un mot, ou un groupe de mots) catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates, etc.

## Identification et categorisation

- **Entités** (personnes, organisations, lieu)
- **Temps** (dates, temps et durée)
- **Quantité** (valeurs monétaires, mesures, pourcentages et nombres cardinaux)

*La décision du député indépendant **Andrew Wilkie** de retirer son soutien au gouvernement **Travailliste** minoritaire a semblé dramatique, mais cela ne devrait pas menacer davantage sa stabilité. Lorsque, après l'élection de **2010**, **Wilkie**, **Rob Oakeshott**, **Tony Windsor** et les **Verts** ont décidé de soutenir le **parti Travailliste**, ils n'ont donné que deux garanties: la confiance et l'approvisionnement.*

# Reconnaissance d'entités nommées

La **reconnaissance d'entités nommées** est une sous-tâche de l'activité d'extraction d'information dans des corpus documentaires. Elle consiste à rechercher des objets textuels (c'est-à-dire un mot, ou un groupe de mots) catégorisables dans des classes telles que noms de personnes, noms d'organisations ou d'entreprises, noms de lieux, quantités, distances, valeurs, dates, etc.

## Identification et **categorisation**

- **Entités** (personnes, organisations, lieu)
- **Temps** (dates, temps et durée)
- **Quantité** (valeurs monétaires, mesures, pourcentages et nombres cardinaux)

*La décision du député indépendant **Andrew Wilkie** de retirer son soutien au gouvernement **Travailliste** minoritaire a semblé dramatique, mais cela ne devrait pas menacer davantage sa stabilité. Lorsque, après l'élection de **2010**, **Wilkie**, **Rob Oakeshott**, **Tony Windsor** et les **Verts** ont décidé de soutenir le **parti Travailliste**, ils n'ont donné que deux garanties: la confiance et l'approvisionnement.*

# NER: Evaluation

	Correct	Not correct
Selected	TP	FP
Not selected	FN	TN

**Précision:** % d'elements pertinents retrouvés rapporté au nombre d'elements total proposé par le systeme.

**Rappel:** % d'elements pertinents retrouvés au regard du nombre d'elements pertinents que possède le jeu de donnees

**F-measure:** weighted harmonic mean

# Precision/Rappel/F1 pour IE/NER

- Le rappel et la précision sont simples à calculer pour des tâches telles que la catégorisation des textes, où il n'y a qu'une seule taille des éléments (documents)
- La mesure se comporte un peu drôlement pour IE / NER quand il y a des erreurs de limites (qui sont communes):

First Bank of Chicago announced earnings ...

- Cela compte à la fois comme un fp et un fn
- Ne rien sélectionner aurait été mieux
- Certains autres métriques (par exemple, le MUC score) donnent un crédit partiel (selon des règles complexes)



# 3 approches standard pour NER (et IE)

1. Expressions régulières écrites à la main
2. Classifieurs (approches supervisées)
3. Modèle de Markov

# Patrons écrits à la main pour NER

Si on doit extraire à partir de pages Web générées automatiquement, les modèles de regex simples fonctionnent (généralement).

## Page Amazon

```
<div class="buying"><h1 class="parseasinTitle"><span  
id="btAsinTitle" style="">(.*?)</span></h1>
```

Pour certains types d'entités restreints et communs dans du texte non structuré, des modèles de regex simples fonctionnent également (généralement).

Trouver numéros de téléphone aux Etats Unis :

```
(?:\((?[0-9]{3}\)?)?[-.]?[0-9]{3}[-.]?[0-9]{4}
```

# Natural Language Processing-based Hand-written Information Extraction

For unstructured human-written text, some NLP may help

- **Part-of-speech (POS) tagging**
  - Mark each word as a noun, verb, preposition, etc.
- **Syntactic parsing**
  - Identify phrases: NP, VP, PP
- **Semantic word categories** (e.g. from WordNet)
  - KILL: kill, murder, assassinate, strangle, suffocate
- Cascaded regular expressions to match relations
  - Higher-level regular expressions can use categories matched by lower-level expressions

# Rules-based extracted examples

Determining which person holds what office in what organization

**[person] , [office] of [org]**

Vuk Draskovic, leader of the Serbian Renewal Movement

**[org] (named, appointed, etc.) [person] Prep [office]**

NATO appointed Wesley Clark as Commander in Chief

Determining where an organization is located

**[org] in [loc]**

NATO headquarters in Brussels

**[org] [loc] (division, branch, headquarters, etc.)**

KFOR Kosovo headquarters

# The ML sequence model approach to NER

## Training

1. Collect a set of representative training documents
2. Label each token for its entity class or other (O)
3. Design feature extractors appropriate to the text and classes
4. Train a sequence classifier to predict the labels from the data

## Testing

1. Receive a set of testing documents
2. Run sequence model inference to label each token
3. Appropriately output the recognized entities

# The full task of Information Extraction

As a family of techniques:

Information Extraction =  
segmentation + classification + association + clustering

For years, [Microsoft Corporation](#) [CEO Bill Gates](#) railed against the economic philosophy of open-source software with Orwellian fervor, denouncing its communal licensing as a "cancer" that stifled technological innovation.

Now [Gates](#) himself says [Microsoft](#) will gladly disclose its crown jewels--the coveted code behind the Windows operating system--to select customers.

"We can be open source. We love the concept of shared source," said [Bill Veghte](#), a [Microsoft VP](#). "That's a super-important shift for us in terms of code access."

[Richard Stallman](#), [founder](#) of the [Free Software Foundation](#), countered saying...

[Microsoft Corporation](#)

[CEO](#)

[Bill Gates](#)

[Gates](#)

[Microsoft](#)

[Bill Veghte](#)

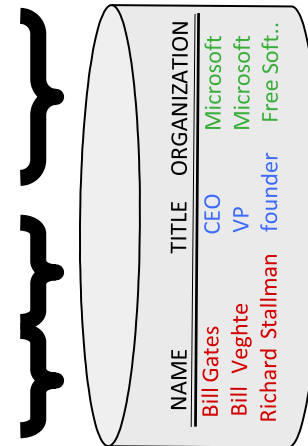
[Microsoft](#)

[VP](#)

[Richard Stallman](#)

[founder](#)

[Free Software Foundation](#)



# Arity of relations

Jack Welch will retire as CEO of General Electric tomorrow. The top role at the Connecticut company will be filled by Jeffrey Immelt.

## Single entity

*Person:* Jack Welch

*Person:* Jeffrey Immelt

*Location:* Connecticut

## Binary relationship

*Relation:* Person-Title

*Person:* Jack Welch

*Title:* CEO

*Relation:* Company-Location

*Company:* General Electric

*Location:* Connecticut

## N-ary record

*Relation:* Succession

*Company:* General Electric

*Title:* CEO

*Out:* Jack Welch

*In:* Jeffrey Immelt

# Association task: Relation Extraction

Checking if groupings of entities are **instances of a relation**

## 1. Manually engineered rules

Rules defined over words/entities:

`<company> located in <location>`

Rules defined over parsed text:

`((Obj <company>) (Verb located) (*) (Subj <location>))`

## 1. Machine Learning-based

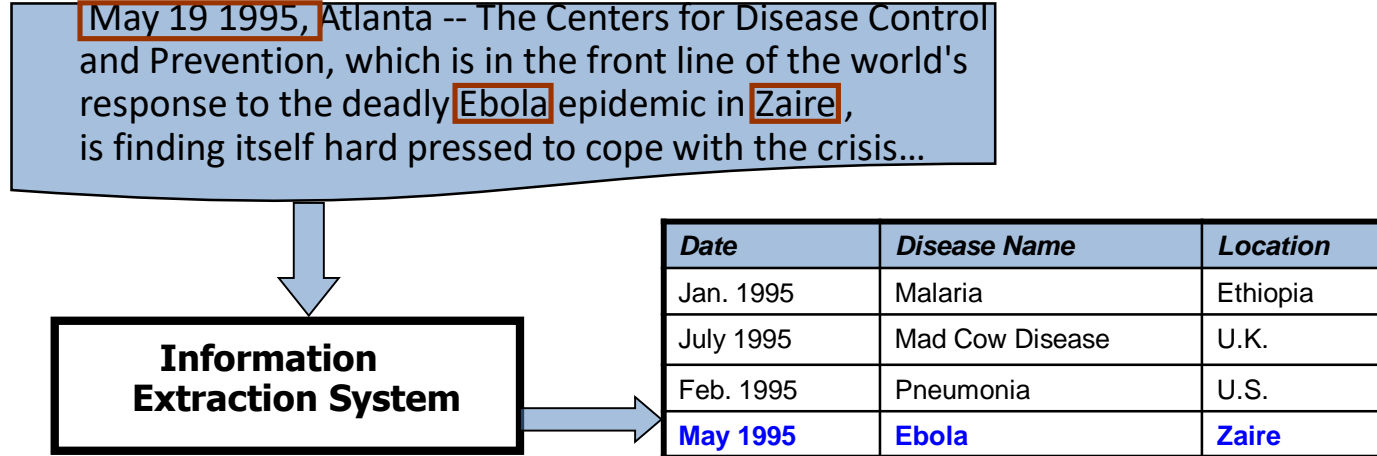
**Supervised:** Learn relation classifier from examples

**Partially-supervised:** bootstrap rules/patterns from “seed” examples





# Example



# Why Relation Extraction?

- Create new **structured knowledge bases**, useful for any application
- Augment current knowledge bases
  - Adding words to WordNet thesaurus, facts to FreeBase or DBpedia
  - Support question answering

*The granddaughter of which actor starred in the movie “E.T.”?*

*(acted-in ?x “E.T.”)(is-a ?y actor)(granddaughter-of ?x ?y)!*

# How to build relation extractor

1. Hand-written patterns
2. Supervised machine learning
3. Semi-supervised and unsupervised
  - Bootstrapping (using seeds)
  - Distant supervision
  - Unsupervised learning from the web

# Hand written patterns

"Agar is a substance prepared from a mixture of **red algae, such as Gelidium,** for laboratory or industrial use"



What does *Gelidium* mean?  
How do you know?

## Patterns for extracting IS-A relation (hyponyms)

"Y such as X ((, X)\* (, and|or) X)"!

"such Y as X"!

"X or other Y"!

"X and other Y"!

"Y including X"!

"Y, especially X"!

# (Hearst 1992)'s patterns for extracting IS-A relations

Hearst pattern	Example occurrences
X and other Y	...temples, treasures, <b>and other</b> important civic buildings.
X or other Y	Bruises, wounds, broken bones <b>or other</b> injuries...
Y such as X	The bow lute, <b>such as</b> the Bambara ndang...
Such Y as X	... <b>such</b> authors <b>as</b> Herrick, Goldsmith, and Shakespeare.
Y including X	...common-law countries, <b>including</b> Canada and England...
Y, especially X	European countries, <b>especially</b> France, England, and Spain...

# Extracting richer relations using rules

**Intuition:** relations often hold between specific entities

`located-in (ORGANIZATION, LOCATION)`

`founded (PERSON, ORGANIZATION)`

`cures (DRUG, DISEASE)`

Start with Named Entity tags to help extract relation!



# NE aren't quite enough. Which relations hold between two entities?



Drug

Cure?  
Prevent?  
Cause?



Disease



# NE aren't quite enough. Which relations hold between two entities?



PERSON

Founder?

Investor?

Member?

Employee?

President?



ORGANIZATION



# Extracting richer relations using rules and NE

Who holds what office in what organization?

**PERSON**, **POSITION** of **ORG**

- **George Marshall**, **Secretary of State** of **the United States**

**PERSON** (named|appointed|chose|etc.) **PERSON** Prep? **POSITION**

- **Truman** appointed **Marshall** **Secretary of State**

**PERSON** [be]? (named|appointed|etc.) Prep? **ORG** **POSITION**

- **George Marshall** was named **US** **Secretary of State**

# Hand-built patterns for relations



Human patterns tend to be high-precision

- Can be tailored to specific domains



Human patterns are often low-recall

- A lot of work to think of all possible patterns!
- Don't want to have to do this for every relation!
- We'd like better accuracy

# Supervised Machine Learning

- Choose a set of **relations** we'd like to extract
- Choose a set of relevant **named entities**
- Find and **label data**
- Choose a representative corpus
- Label the Named Entities in the corpus
- Hand-label the relations between these entities
- Break into training, development, and test
- **Train a classifier** on the training set

# Supervised Relation Extraction



Can get high accuracies with enough hand-labeled training data, if test similar enough to training



Labeling a large training set is expensive  
Supervised models are brittle, don't generalize well to different genres

# Semi-supervised and unsupervised

**Bootstrapping:** use the seeds to directly learn to populate a relation

Gather a set of **seed pairs** that have relation R

**Iterate:**

1. Find sentences with these pairs
2. Look at the context between or around the pair and generalize the context to create patterns
3. Use the patterns for grep for more pairs

# Bootstrapping

`<Mark Twain, Elmira>` Seed tuple

- Grep (google) for the environments of the seed tuple

“Mark Twain is buried in Elmira, NY.”

**X is buried in Y**

“The grave of Mark Twain is in Elmira”

**The grave of X is in Y**

“Elmira is Mark Twain’s final resting place”

**Y is X’s final resting place.**

- Use those patterns to grep for new tuples
- Iterate

# Resolving coreference (both within and across documents)

John Fitzgerald Kennedy was born at 83 Beals Street in Brookline, Massachusetts on Tue 29, 1917, at 3:00 pm,[7] the second son of Joseph P. Kennedy, Sr., and Rose Fitzgerald; Return, was the eldest child of John "Honey Fitz" Fitzgerald, a prominent Boston political figure who was the city's mayor and a three-term member of Congress. Kennedy lived in Brookline for 10 years and attended Edward Devotion School, Noble and Greenough Lower School, and the Boston Latin School, through 4th grade. In 1927, the family moved to 5040 Independence Avenue in the Bronx, New York City; two years later, they moved to 294 Pondfield Road in Bronxville, New York, where Kennedy was a member of Scout Troop 2 (and was the first Boy Scout to become President).[8] Kennedy spent summers with his family at their home in Hyannisport, Massachusetts, and Christmas and Easter holidays with his family at their winter home in Palm Beach, Florida. For the 5th through 7th grade, Kennedy attended Riverdale Country School, a private school for boys. For 8th grade in September 1930, the 13-year old Kennedy attended Canterbury School in New Milford, Connecticut.



# Rough Accuracy of Information Extraction

Information type	Accuracy
Entities	90-98%
Attributes	80%
Relations	60-70%
Events	50-60%

**Errors cascade** (error in entity tag = > error in relation extraction)

These are very rough, actually optimistic, numbers

Hold for well-established tasks, but lower for many specific/novel IE tasks



# Let's go back to NEs...

German foreign minister **Westerwelle** visits **Ghana**.

**William Hague** and **Angelina Jolie** visit **Eastern DRC**.

**Blackstone Group LP** (BX) agreed to buy 23 industrial properties in **southern Virginia** and the **Washington** and **Baltimore** metropolitan areas from **First Potomac Realty Trust** (FPO) for \$241.5 million.

<input type="checkbox"/>	FirstPerson
<input type="checkbox"/>	JobTitle
<input checked="" type="checkbox"/>	Location
<input type="checkbox"/>	Lookup
<input type="checkbox"/>	Money
<input checked="" type="checkbox"/>	Organization
<input checked="" type="checkbox"/>	Person

- We know the type of the Named Entity but nothing more...
- What kind of organization is the **Blackstone Group LP**?
- What is the job of **William Hague**?
- Where is **Eastern DRC** and what **DRC** stands for?

=> ONLY SEMANTICS: annotation type name

# Need more semantics:

- **To coreference** DRC with “Democratic Republic of Congo”
- **To avoid scattered knowledge:** cities are locations, cities have zip code, etc.
- **To disambiguate:** which Washington (state/city)?
- To use **extracted information to allow for queries** as:
  - European politicians who visited an African country?
  - Politicians and actors travelling together?
- To use **extracted information to add information to our own database/knowledge base**

# Semantic search in Google



## Paris - Wikipedia, the free encyclopedia

<https://en.wikipedia.org/wiki/Paris> - Wikipedia

Paris was founded in the 3rd century BC by a Celtic people called the Parisii, who gave the city its name. By the 12th century, Paris was the largest city in the ...  
Eiffel Tower - History of Paris - Ile-de-France - Parisii

## Paris Hilton (@ParisHilton) | Twitter

<https://twitter.com/ParisHilton>

19 hours ago

#HalloweenMakeUpTransformation  
contest is over tomorrow 10AM PST  
#LastChance #KillIt [ow.ly/dKKAjr](https://www.youtube.com/watch?v=dKKAjr)

22 hours ago

At the launch event for my new  
#ParisHiltonCosmetics Line in  
Shanghai. Love my new lipgloss, such  
a... [instagram.com/p/8-54\\_kqgN...](https://www.instagram.com/p/8-54_kqgN...)

## Images for paris

Report images



[More images for paris](#)

## Paris Tourism: Best of Paris, France - TripAdvisor

[www.tripadvisor.com/Tourism-g187147-Paris\\_Ile\\_de\\_Franc...](http://www.tripadvisor.com/Tourism-g187147-Paris_Ile_de_Franc...) - TripAdvisor

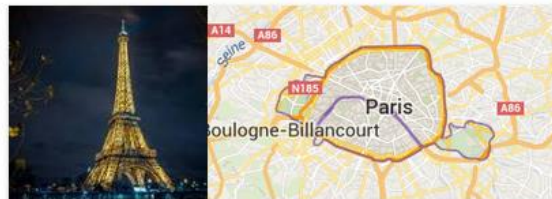
Paris Tourism: TripAdvisor has 1932824 reviews of Paris Hotels, Attractions, and Restaurants making it your best Paris resource.

[Things to do in Paris](#) - [Paris Hotels](#) - [Restaurants](#) - [Paris Travel Forum](#)

## Paris, France - Lonely Planet

[www.lonelyplanet.com](http://www.lonelyplanet.com) > Europe > Western Europe > France - Lonely Planet

Paris off the beaten path. Information on when and where to go, activities, history,



## Paris

Capital of France

Paris, France's capital, is a major European city and a global center for art, fashion, gastronomy and culture. Its picturesque 19th-century cityscape is crisscrossed by wide boulevards and the River Seine. Beyond such landmarks as the Eiffel Tower and the 12th-century, Gothic Notre-Dame cathedral, the city is known for its cafe culture, and designer boutiques along the Rue du Faubourg Saint-Honore.

**Area:** 40.7 mi<sup>2</sup>

**Weather:** 54°F (12°C), Wind N at 2 mph (3 km/h), 68% Humidity

**Local time:** Monday 2:22 PM

**Hotels:** 3-star averaging \$110, 5-star averaging \$360. [View hotels](#)

**Getting there:** 1 h 25 min flight, around \$105. [View flights](#)

**Population:** 2.244 million (2010) UNdata

## Points of interest

[View 10+ more](#)



Disneyland  
Paris



Eiffel Tower



Notre  
Dame de  
Paris



The Louvre



Arc de  
Triomphe

# Searching for “Things” not “strings”



Paris - Wikipedia, the free encyclopedia

<https://en.wikipedia.org/wiki/Paris> - Wikipedia

Paris was founded in the 3rd century BC by a Celtic people called the Parisii, who gave the city its name. By the 12th century, Paris was the largest city in the ...  
Eiffel Tower - History of Paris - Ile-de-France - Parisii

Paris Hilton (@ParisHilton) | Twitter

<https://twitter.com/ParisHilton>

10 hours ago

22 hours ago

.500 millions entities that Google “knows” about  
.Used to provide more accurate search results  
.Summary of the information of the entity you  
.are searching about

Paris Tourism: Best of Paris, France - TripAdvisor

[www.tripadvisor.com/Tourism-g187147-Paris\\_Ile\\_de\\_Franc...](http://www.tripadvisor.com/Tourism-g187147-Paris_Ile_de_Franc...) - TripAdvisor

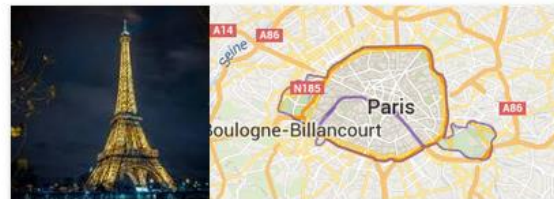
Paris Tourism: TripAdvisor has 1932824 reviews of Paris Hotels, Attractions, and Restaurants making it your best Paris resource.

Things to do in Paris - Paris Hotels - Restaurants - Paris Travel Forum

Paris, France - Lonely Planet

[www.lonelyplanet.com](http://www.lonelyplanet.com) > Europe > Western Europe > France - Lonely Planet

Paris off the beaten path. Information on when and where to go, activities, history,



## Paris

France

France's capital, is a major European city and a global center of fashion, gastronomy and culture. Its picturesque 19th-century center is crisscrossed by wide boulevards and the River Seine. Other landmarks include the Eiffel Tower and the 12th-century, Notre-Dame cathedral, the city is known for its cafe culture, and boutiques along the Rue du Faubourg Saint-Honoré.

7 mi²

54°F (12°C), Wind N at 2 mph (3 km/h), 68% Humidity

Monday 2:22 PM

5-star averaging \$110, 5-star averaging \$360. [View hotels](#)

here: 1 h 25 min flight, around \$105. [View flights](#)

on: 2.244 million (2010) UNdata

## Points of interest

[View 10+ more](#)



Disneyland Paris



Eiffel Tower



Notre Dame de Paris



The Louvre



Arc de Triomphe

# Facebook graph search

The screenshot displays a Facebook search results page for the query "Current Tesco employees who like Horses". The interface is divided into three main sections: a list of search results on the left, a refinement sidebar on the right, and an extension section at the bottom right.

**Search Results (Left):** The results are presented as a list of profile cards. Each card shows a blurred profile picture, the user's name, and their current employer (Tesco). Below the name, there are icons and text indicating interests (e.g., "Likes Horses and Dogs"), education (e.g., "Studied at"), location (e.g., "Lives in Liverpool"), and music preferences (e.g., "Listens to"). At the bottom of each card are buttons for "Add Friend", "Message", and a search icon.

**Refinement Sidebar (Right):** This section is titled "More Than 100 People" and includes a "View Grid" link. It contains a "REFINE THIS SEARCH" section with various filters: Gender, Relationship, Current Employer (set to "Tesco"), Position, Employer Location, Time Period, Current City, Hometown, School, Friendship, and Likes (set to "Horses"). A "SEE MORE" link is located below these filters.

**Extend This Search (Bottom Right):** This section is titled "EXTEND THIS SEARCH" and includes three options: "More pages they like" (with a Tesco logo), "Photos of these people" (with a photo of a person), and "These people's friends" (with a photo of a person).

# Semantic Annotation

**Semantic:** link the annotation to a concept in an ontology.

- The semantic link **connects the text mention to knowledge about the concept that is mentioned.**
- The mention can link to an instance, a class, or a property – i.e. to a **resource**
- **Use the semantic link to access additional data about the concept** – use for disambiguation and further annotation processing
- Use for NER, IE, querying, ...

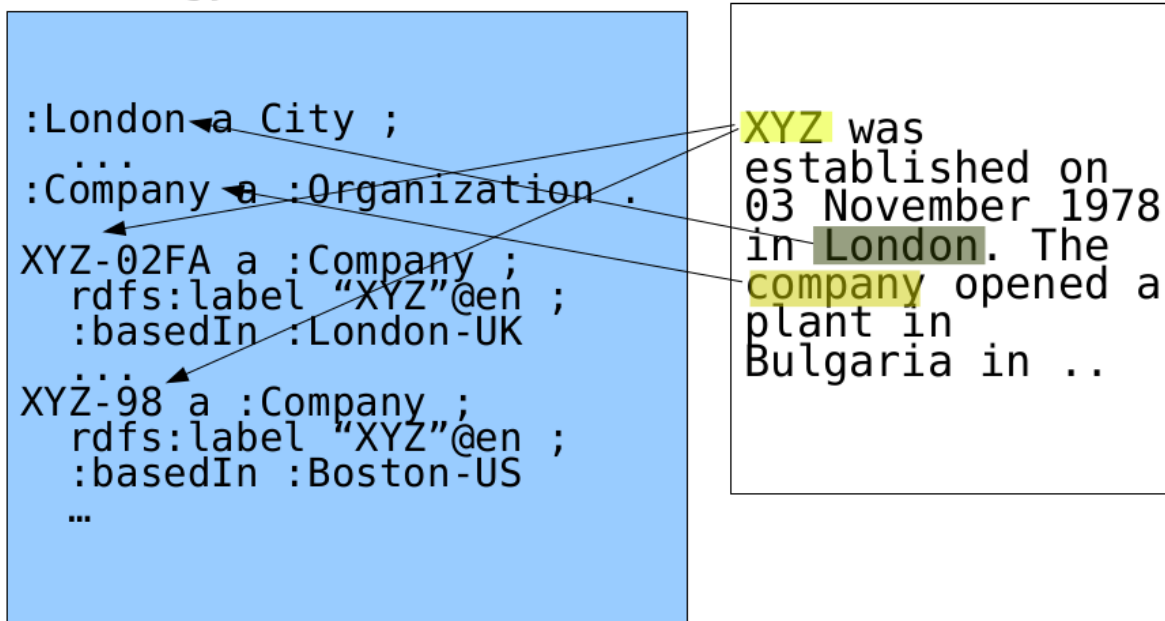
# Semantic Annotation

## Ontology

```
:London a City ;  
:Company a :Organization .  
XYZ-02FA a :Company ;  
  rdfs:label "XYZ"@en ;  
  :basedIn :London-UK  
XYZ-98 a :Company ;  
  rdfs:label "XYZ"@en ;  
  :basedIn :Boston-US  
...
```

## Document

XYZ was  
established on  
03 November 1978  
in London. The  
company opened a  
plant in  
Bulgaria in ..



# Semantic Annotation

## Ontology

```
:London a City ;  
:Company a :Organization .  
XYZ-02FA a :Company ;  
  rdfs:label "XYZ"@en ;  
  :basedIn :London-UK  
XYZ-98 a :Company ;  
  rdfs:label "XYZ"@en ;  
  :basedIn :Boston-US  
...
```

## Document

XYZ was  
established on  
03 November 1978  
in London. The  
company opened a  
plant in  
Bulgaria in ..



# Semantic Annotation vs “traditional”

- Link to hierarchy of concepts instead of flat set of concepts
- Larger space of possible annotations
- - harder to get it right
- + candidate concepts have associated knowledge that can be used to support decision
- + found concepts can be generalized based on ontology:  
context(company) < context(organization)

# Entity linking

article discussion edit this page history

## Plant

From Wikipedia, the free encyclopedia

*For other uses, see Plant (disambiguation).*

**Plants** are a major group of living things including familiar organisms such as trees, flowers, herbs, ferns, and mosses. About 350,000 species of plants, defined as seed plants, bryophytes, ferns and fern allies, have been estimated to exist. As of 2004, some 287,655 species had been identified, of which 258,650 are flowering and 15,000 bryophytes.

**Tree**

From Wikipedia, the free encyclopedia

*For other senses of this word, see tree (disambiguation).*

A **tree** is a large, perennial, woody plant. Though there is no set definition regarding minimum size, the term generally applies to plants at least 6 m (20 ft) high at maturity and, more importantly, having

**Fossil range: Middle-Late Ordovician - Recent**

**Species**

From Wikipedia, the free encyclopedia

*This article is about biology. For the movie, see Species.*

In biology, a **species** is one of the basic units of biodiversity classification; a species is assigned a two-part name, the **genus** is listed first (with its leading letter capitalized), followed by the **specific epithet**. For example, humans belong to the genus *Homo*, and species *Homo sapiens*. The name of the species is the whole name, and the second term (which may be called **specific epithet**) is the **specific name**.

# Entity linking?

- NE normalization / canonicalization / sense disambiguation
- DB record linkage / schema mapping (not the focus here)
- Knowledge base population
- Entity linking
- Wikification
- Semantic linking

# Entity linking: main problem

## Linking free text to entities

### Any piece of text

- news documents
- blog posts
- tweets
- queries
- ...



### **Entities:** typically taken from a knowledge base

- Wikipedia
- Freebase
- ...

# Common steps


1. Determine “linkable” phrases
  - **mention detection – MD**
2. Rank/Select candidate entity links
  - **link generation – LG**
  - may include NILs (null values, i.e., no target in KB)
3. (Use “context” to disambiguate/filter/improve)
  - **disambiguation – DA**

Input Text

 Italiano  English

momentum) degeneracy is removed due to a geometric gradient onto a metasurface. The alliance of spin optics and metamaterials offers the dispersion engineering of a structured matter in a polarization helicity-dependent manner. We show that polarization-controlled optical modes of metamaterials arise where the spatial inversion symmetry is violated. The emerged spin-split dispersion of spontaneous emission originates from the spin-orbit interaction of light, generating a selection rule based on symmetry restrictions in a spin-optical metamaterial. The inversion asymmetric metasurface is obtained via anisotropic optical antenna patterns. This type of metamaterial provides a route for spin-controlled nanophotonic applications based on the design of the metasurface symmetry properties.

Many links



Few links

Reset

TAGME!

Tagged text

Topics

Spin optics provides a route to control light, whereby the photon helicity (spin angular momentum) degeneracy is removed due to a geometric gradient onto a metasurface. The alliance of sp matter in a p optical mode emerged spin interaction of optical metar

**Degenerate energy levels**

In physics, two or more different quantum states are said to be degenerate if they are all at the same energy level. Statistically this means that they are all equally probable of being filled, and in...

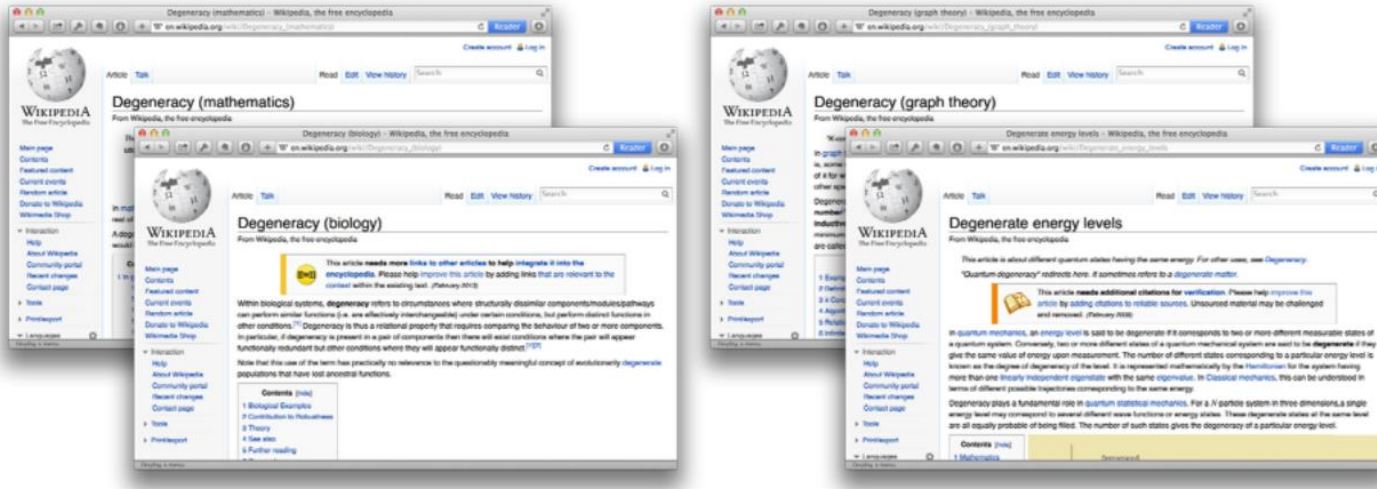
...e dispersion engineering of a structured ...r. We show that polarization-controlled spatial inversion symmetry is violated. The emission originates from the spin-orbit ed on symmetry restrictions in a spin- ed via anisotropic

# Mention detection (MD)

Q ... degeneracy is removed ...

# Link Generation (LG)

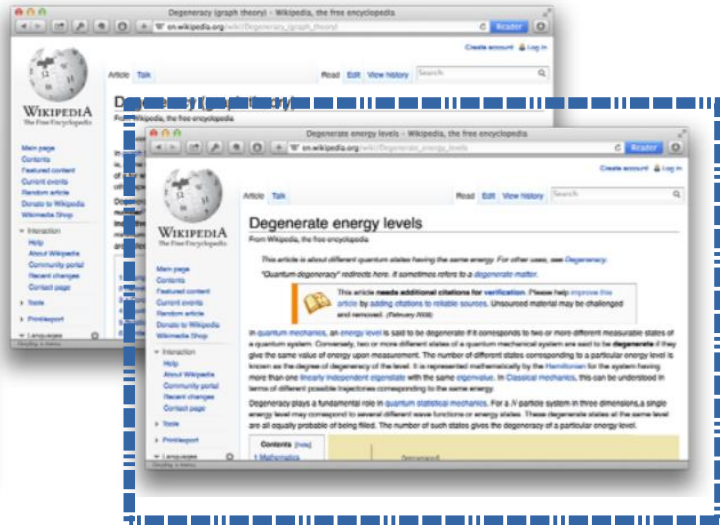
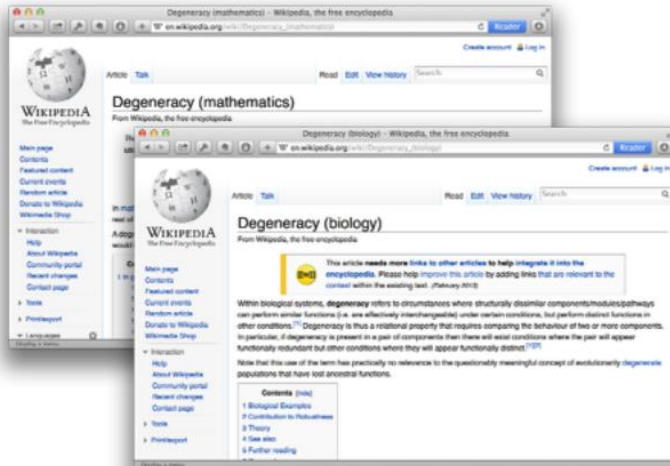
Q ... degeneracy ...





# Disambiguation (DA)

Q ... degeneracy ...





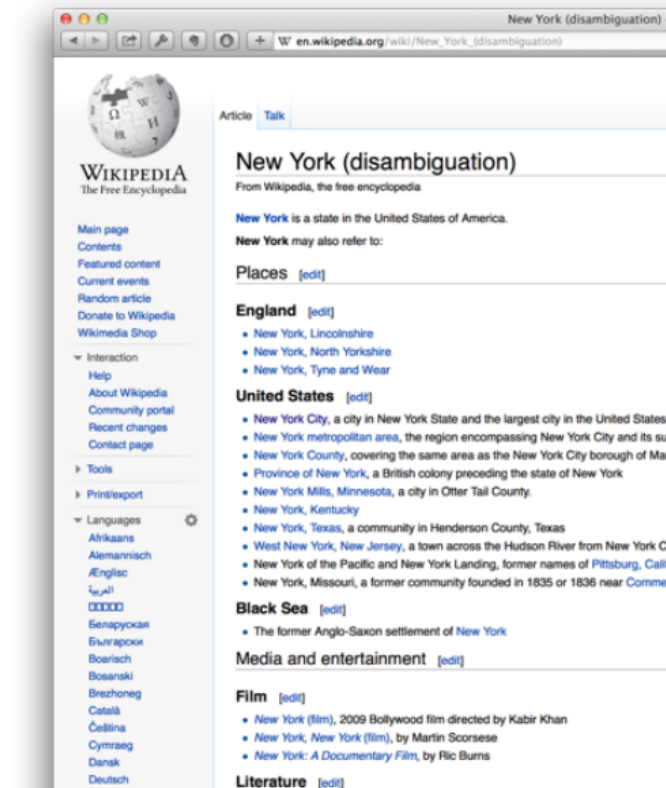
# Wikipedia

- Basic element: article (proper)
- But also
  - redirect pages
  - disambiguation pages
  - category/template pages
  - admin pages
- Hyperlinks
  - use “unique identifiers” (URLs)
    - [[United States]] or [[United States|American]]
    - [[United States (TV series)]] or [[United States (TV series)|TV show]]



# Wikipedia

- Senses of an ambiguous phrase
- Short description
- (Possible) categorization
- Non-exhaustive



# Wikipedia-based methods

- keyphraseness(w) [Mihalcea & Csomai 2007]

$$\frac{CF(w_l)}{CF(w)} \rightarrow \begin{array}{l} \text{Collection frequency} \\ \text{term } w \text{ as a link to another} \\ \text{Wikipedia article} \end{array}$$

↓

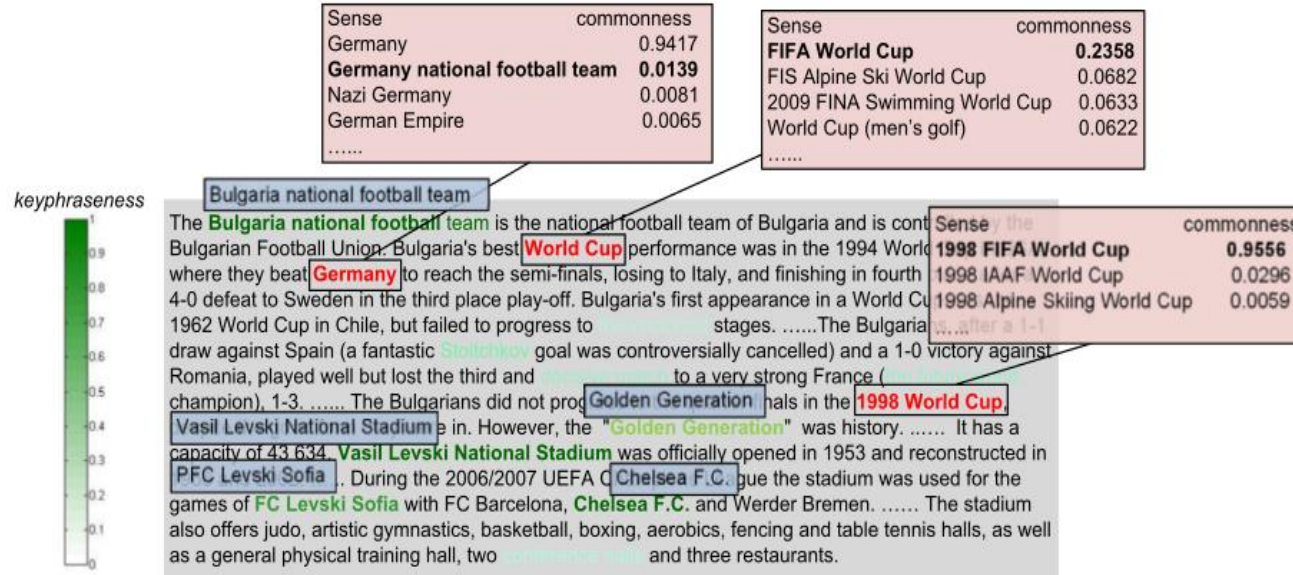
**Collection frequency**  
term w

- commonness(w,c) [Medelyan et al. 2008]

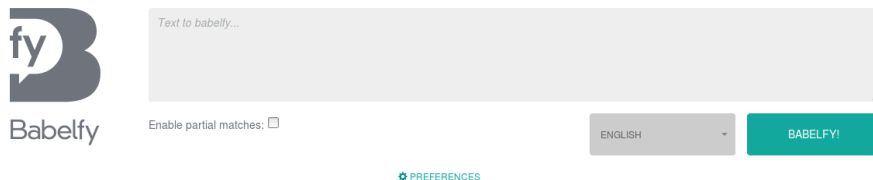
$$\frac{|L_{w,c}|}{\sum_{c'} |L_{w,c'}|}$$

⏟  
**Number of links**  
with target c' and anchor text w

# Wikipedia-based methods



# An example: BABELFY



- unified, multilingual, graph-based approach to **Entity Linking and Word Sense Disambiguation**
- based on the multilingual semantic network BabelNet

BABELFY performs disambiguation and entity linking in three steps:

- i) It associates with each vertex of the BabelNet semantic network, i.e., either concept or named entity, a semantic signature, that is, a set of related vertices.
- ii) Given an input text, it extracts all the linkable fragments from this text and, for each of them, lists the possible meanings according to the semantic network.
- lii) It creates a graph-based semantic interpretation of the whole text by linking the candidate meanings of the extracted fragments using the previously-computed semantic signatures. It then extracts a dense subgraph of this representation and selects the best candidate meaning for each fragment.

# Ontology population

- Annotate document and find mentions of what could be (new) instances in the ontology
  - Use traditional NER, linked to ontology
  - Use semantic annotation based on existing knowledge
- Use ML
- Create ontology instances and property values
- (“ABOX”) from the final annotations

# Ontology population

```
:London a City ;
:Company a :Organization .
```

XYZ was  
established on  
03 November 1978  
in London. The  
company opened a  
plant in  
Bulgaria in ..

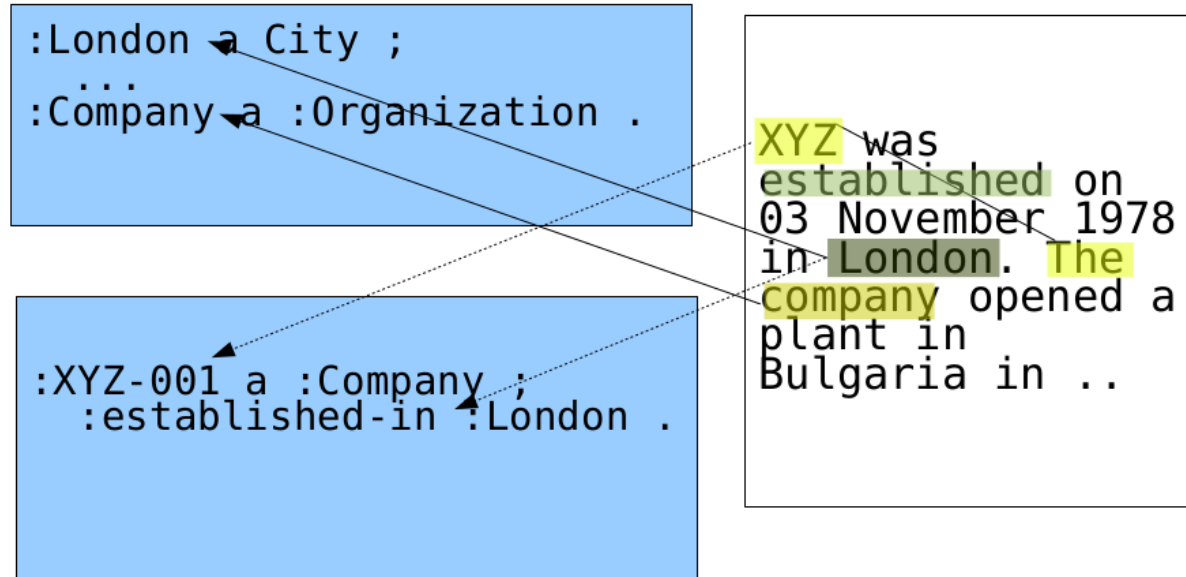


# Ontology population

```
:London a City ;
:Company a :Organization .
```

XYZ was  
established on  
03 November 1978  
in London. The  
company opened a  
plant in  
Bulgaria in ..

# Ontology population



# Ontology population

- Populate Ontology with Instances:
  - Of classes
  - Of properties connecting class instances with other class instances or values (literals)
  - Graph describing n-ary relations or events ...
- Strategy
  - Place in domain ontology?
  - Place in intermediate ontology/KB?

# Ontology population

- Place directly in domain ontology:
  - + Simple & straight-forward
  - Cannot model likelihoods, hard to model meta information (where from, which context)
  - Can easily leave sub-language or become inconsistent
  - Knowledge arrives incrementally but has dependencies
- Place in intermediate ontology
- - Processing more complex
- Appropriate model for intermediate ontology?
- + Can do iterative improvement
- + Can model meta information

# Recap

## . Semantic Annotation

- Mentions of instances in the text are annotated wrt concepts (classes) in the ontology.
- Requires that instances are disambiguated.
- It is the **text** which is modified.

## . Ontology Population

- Generates new instances in an ontology from a text.
- Links unique mentions of instances in the text to instances of concepts in the ontology.
- It is the **ontology** which is modified.