Named Entity Processing for Historical Texts

DEFINITION, RESOURCES, APPROACHES, APPLICATIONS

Maud Ehrmann [1] **Matteo Romanello** [1] **Simon Clematide** [2]

9 July 2019

Slides `https://github.com/impresso/named-entity-tutorial-dh2019`

[1]EPFL-DHLAB, Lausanne, Switzerland

[2]Institute for Computational Linguistics, University of Zurich, Switzerland

Schedule

- **09h00-10h30**: Theoretical part
- **10h30-11h00**: Coffee break
- **11h00-13h00**: Theoretical part /Hands on
- **13h00-14h00**: Lunch break
- **14h00-14h45**: Hands-on
- **14h45-15h15**: Coffee break
- **15h15-16h00**: Hands-on

Who are you?

Interactive session

What is your background? In which type of institution do you work?
TODO: decide if we use mentimeter.com to do interactive questions.

Introduction to named entity processing: Outline

1. Context and Applications
2. Definition
3. Resources
4. Recognition et classification
5. Linking (quickly)
6. Evaluation

1. Context and Applications

1. Context and Applications

1.1 Introduction

Context


The Economist
The DPJ messes up
The economic shift from West to East
Naxalite violence spreads in India
Genetically modified crops blossom
The right to eat cats and dogs
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

Data

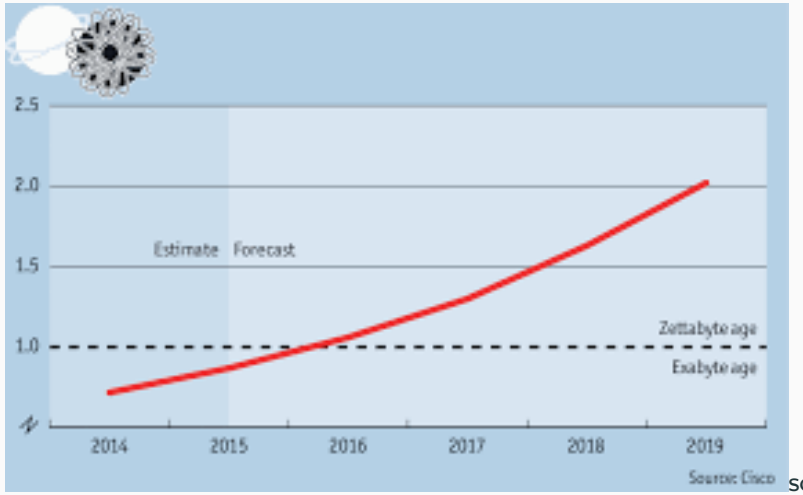
What: EVERYTHING

i.e. text, images and audio material published on news websites, social medias, collaborative platforms, smartphones, sensors, etc.

Profil (1/2)

How much: astronomical increase

- quantity: doubles every 2 years
- traffic: entered the zettabyte era in (1 trillion gigabytes)
- storage: projection of 44 zettabytes in 2020



source: `http://www.theworldin.com/article/12107/charting-change`

Profil (2/2)

Nature:
80 to 90% of data are non structured, i.e. without pre-defined model nor format.

Challenges:

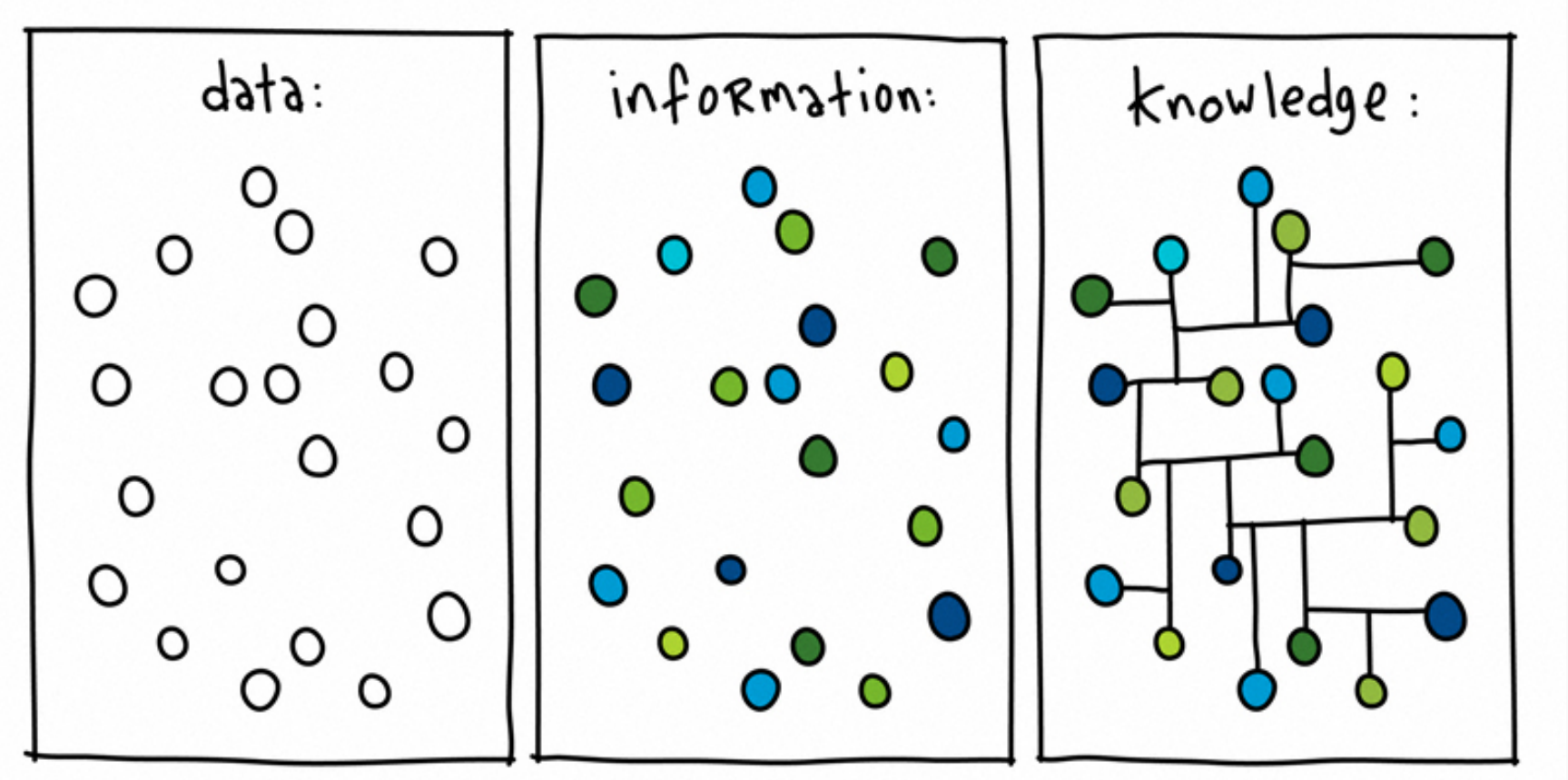
- storage more and more expensive
- above all: data exploitation, extract useful information

Clarification

Data	Information	Knowledge
basic description of a reality	data with a meaning constructing a representation of reality	information with a truth
temperature measurement	*plot on the evolution of the average minima and maxima in a given place, by month*	*fact that the temperature on earth is increasing because of human activity*
series of journalistic articles	*name of people and their polarities*	*opinion of the media vis--vis personalities*

inspired from: `http://www.college-de-france.fr/site/serge-abiteboul/_inaugural-lecture.htm`

Clarification



Source: gapingvoid - Culture Design Group

Semi-structured data

Cannes Film Festival

From Wikipedia, the free encyclopedia

Coordinates: 43°33′03.10″N 7°01′02.10″E

The **Cannes Festival** (/kæn/) (French: ***Festival de Cannes***), named until 2002 as the **International Film Festival** (*Festival international du film*) and known in English as the **Cannes Film Festival**, is an annual film festival held in Cannes, France, which previews new films of all genres, including documentaries, from all around the world. Founded in 1946, the invitation-only festival is held annually (usually in May) at the Palais des Festivals et des Congrès.[1][2][3]

On 1 July 2014, co-founder and former head of French pay-TV operator Canal+ Pierre Lescure took over as President of the festival. The Board of Directors also appointed Gilles Jacob as Honorary President of the festival.[4][5][6]

The 2016 Cannes Film Festival took place between 11 and 22 May 2016. Australian film director George Miller was the President of the Jury. *I, Daniel Blake*, directed by British director Ken Loach, won the Palme d'Or.

In 2017, The Festival de Cannes will celebrate its 70th anniversary edition from May 17 to 28.

Contents [hide]

Cannes Film Festival

FESTIVAL DE CANNES

Location	Cannes, France
Founded	September 20, 1946
Awards	Palme d'Or, Grand Prix
Website	festival-cannes.com

Festival de Cannes ✓
@Festival_Cannes

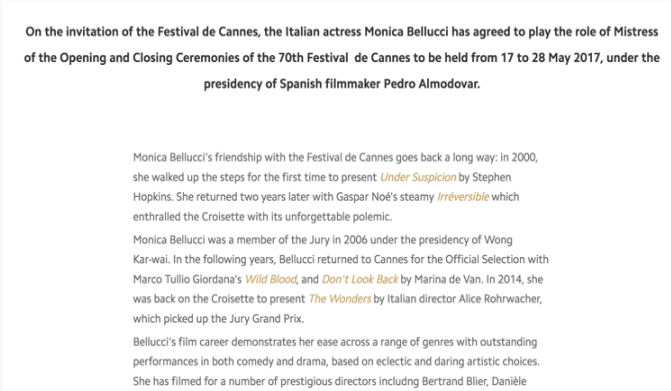
Follow

In French theaters today, testimonies from Ugandan ex-child soldiers : Wrong Elements by Jonathan Littell #SpecialScreening in #Cannes2016

But most of the time,
information is ‘hidden’ in texts

Unstructured data

"On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the opening and Closing Ceremonies of the 70the festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar."



On the invitation of the Festival de Cannes, the Italian actress Monica Bellucci has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th Festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar.

Monica Bellucci's friendship with the Festival de Cannes goes back a long way: in 2000, she walked up the steps for the first time to present *Under Suspicion* by Stephen Hopkins. She returned two years later with Gaspar Noé's steamy *Irréversible* which enthralled the Croisette with its unforgettable polemic.

Monica Bellucci was a member of the Jury in 2006 under the presidency of Wong Kar-wai. In the following years, Bellucci returned to Cannes for the Official Selection with Marco Tullio Giordana's *Wild Blood*, and *Don't Look Back* by Marina de Van. In 2014, she was back on the Croisette to present *The Wonders* by Italian director Alice Rohrwacher, which picked up the Jury Grand Prix.

Bellucci's film career demonstrates her ease across a range of genres with outstanding performances in both comedy and drama, based on eclectic and daring artistic choices. She has filmed for a number of prestigious directors includng Bertrand Blier, Danièle

source: `www.festival-cannes.com`

Information ‘hidden’ in texts

On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar.

PERSON, ORGANIZATION, TIME-EXPR, EVENT

Information Extraction (IE)

The goal is to extract structured information from unstructured texts, ie:

- identifying et categorising information fragments
- linking with knowledge bases
- aggregating to extract further information

Main tasks in IE

- **Named entity processing**:
  recognition, categorization and disambiguation
  - *Monica Belluci* and *Pedro Almodovar* are PERSON.
  - *Monica Belluci* $\xrightarrow{reference}$ `http://dbpedia.org/page/Monica_Bellucci`
- **Temporal expression processing**:
  extraction and normalisation
  - *from 17 to 28 May 2017* is a DURATION
  - *from 17 to 28 May 2017* $\longrightarrow$ [17-05-2017, 28-05-2017]
- **Event extraction**
  - *70th Festival de Cannes* is a FACTUAL, RECURRING EVENT
  - *70th Festival de Cannes* $\xrightarrow{instance\_of}$
    `en.wikipedia.org/wiki/Cannes_Film_Festival`
- **Relation extraction**:
  - *70th Festival de Cannes*, *tookPlace*, [17-05-2017, 28-05-2017]

1. Context and Applications

1.2 A bit of history

From text understanding to information extraction

- **1980s**: objective of automatic text understanding
- A project too ambitious which faces theoretical and technical difficulties:
  - low coverage of grammars
  - too many unresolved ambiguities
  - difficulties in collecting, representing and manipulating knowledge

  $\rightarrow$ generic approach to the comprehension of texts is still a utopia
- **1990s**: decomposition of the task
  - focus on specific elements of interest
  - a template is defined in advance depending on the application
  - local analysis (10-20% of the text is necessary).

The MUC conference series

- *Message Understanding Conference*
- Cycle of 7 evaluation campaigns between 1987 and 1998
- Initiated by the US Office of Naval Research
- Financed by DARPA (Defense Advanced Research Project Agency)

Evolution of MUC conferences

Phase 1: exploratory cycle

- **1987 (MUC-1)** no specific task, military reports on naval operations in telegraphic style;
- **1989 (MUC-2)** definition of predefined templates with 10 champs; definition of evaluation measures (precision and recall).

Evolution of MUC conferences

Phase 2: slot filling of increasingly complex templates

- **1991 (MUC-3)** news reports on terrorist events in Central and South America; form with 18 slots.
- **1992 (MUC-4)** idem, 24 slots
- **1993 (MUC-5)** more complex tasks, tests on out-of-domain material, 2 languages, 11 templates 48 hierarchical slots

MUC 3: Definition of the task of comprehension

Given a document, participants were asked to:

- identify events,
- identify units linked to these events,
- "normalise" these units,
- fill out a descriptive template.

Template MUC-3

19 March - A bomb went off this morning near a power tower in San Salvador leaving a large part of the city without energy, but no casualties have been reported. According to unofficial sources, the bomb - allegedly detonated by urban guerrilla commandos - blew up a power tower in the northwestern part of San Salvador at 0650 (1250 GMT).	
INCIDENT TYPE	bombing
DATE	March 19
LOCATION	El Salvador : San Salvador (city)
PERPETRATOR	urban guerrilla commandos
PHYSICAL TARGET	power tower
HUMAN TARGET	-
EFFECT ON PHYSICAL TARGET	destroyed
EFFECT ON HUMAN TARGET	no injury or death
INSTRUMENT	bomb

Source: Grishman

Evolution of MUC conferences

Phase 3: reformulation of objectives and definition of subtasks

- **1995 (MUC-6)**
  - independent technologies and components
    $\rightarrow$ definition of the sub-task dedicated to "named entities"
  - portable systems
    $\rightarrow$ definition of generic templates
  - consideration of "basics building blocks of understanding"
    $\rightarrow$ co-reference, word sense disambiguation, predicate-argument structure, etc.
- **1997 (MUC-7)** continuation

MUC 6: starting point for work on ENs

- Definition of the notion of named entity
- Definition of the task of recognition and classification of NEs

Afterwards: MET, IREX, CONLL, ACE, ESTER, ETAPE, HAREM, EVALITA, GERMEVAL, TREC, TAC, etc.

Open information extraction

TODO: some slides on this topic

Focus on Digital Humanities

TODO: some slides on this topic: (Matteo/Maud)

- massive digitization
- information also hidden in texts
- no tradition of shared tasks, inherits from what was done in NLP

1. Context and Applications

1.3 NE common definition

Named entities: first definition (NLP)

- Elements "of interest", usually of type *Person*, *Organisation*, *Location*
- Referential units which underlie the meaning of texts

Named entities: different tasks

1. Recognition: detecting, spotting named entities in textual streams (one delimitates NEs 'boundaries' in texts)

2. Classification: categorizing detected segments according to pre-defined semantic categories (one assigns a type)

3. Disambiguation/linking: linking entity mentions to a unique reference (one determines the reference)

4. Relation extraction: discovering relations between NEs (e.g. *father-of*, *born-in*, *alma mater*)

Text

On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar. [...] Monica Bellucci's friendship with the Festival de Cannes goes back a long way: in 2000, she walked up the steps for the first time to present *Under Suspicion* by Stephen Hopkins.

Stanford NER

On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar. [...] Monica Bellucci's friendship with the Festival de Cannes goes back a long way: in 2000, she walked up the steps for the first time to present *Under Suspicion* by Stephen Hopkins.

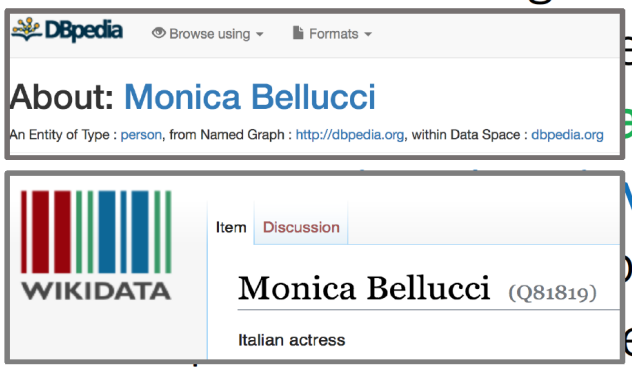
PERSON, ORGANIZATION, LOCATION, DATE

More information?

On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar. [...] Monica Bellucci's friendship with the Festival de Cannes goes back a long way: in 2000, she walked up the steps for the first time to present *Under Suspicion* by Stephen Hopkins.

Linking and relation extraction

On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the emonies of the 70th festival de Cannes to ay 2017, under the presidency of Spanish var. [...] Monica Bellucci's friendship with oes back a long way: in 2000, she walked up e to present *Under Suspicion* by Stephen Hopkins.



director

1. Context and Applications

1.4 Applications

NLP ‘internal’ applications (1/4)

- **Morpho-syntactic tagging**
  - *HyOx,Inc*.
  - *Seat and Porsche has fewer registrations in July 1996.*
- **Syntactic analysis**
  - *He will be replaced by Eliahu Ben-Alissar, a former israeli envoy to Egypt and Jordan.*
  - *He will be replaced by Eliahu Ben-Alissar, a former israeli envoy to Egypt and Likud party.*

NLP 'internal' applications (2/4)

- **Dependency analysis**
  They met in Bagdad. $\rightarrow$ LOCATION*(met, Bagdad)*
- **Coreference**
  John bought a new computer. **It** *was able to train the model.*
- **Translation**
  Jack London was an american writer. London is a busy city.

NLP 'internal' applications (3/4)

- **Word sense disambiguation**
  - *It is difficult to leave Paris on Friday evenings.*
  - *Some wonder if they will leave the Socialist party.*

  What if the meaning of 'leave' ?

NLP 'internal' applications (4/4)

- **Word sense disambiguation**

  It is difficult to leave Paris on Friday evenings.
  → leave = "go away from a place" (#1 WordNet)

  Some wonder if they will leave the Socialist party.
  → leave = "remove oneself from an association with or participation in" (#8 WordNet)

‘Real world’ applications

- **Information extraction and media monitoring**
  - Knowledge base population
  - Alerts on certain topics or entities
- **Cross-lingual document clustering**
  Documents mentioning the same entities are likely to be linked.
- **Summarization**
  NEs are informational ’anchors’ helping to identify key elements of a text

Application in DH

todo (Matteo/Maud)

What about you?

Interactive session

What is/would be your usage of named entity processing?

TODO: decide if we use mentimeter.com to do interactive questions.

context: take-home message

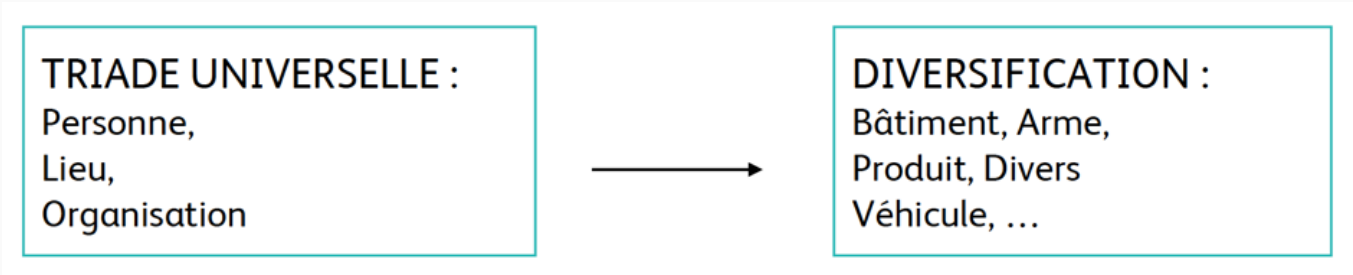
- The concept of NE appeared in the 90s during evaluation campaigns on text understanding.
- NEs quickly gained prominence and became one of the central hubs in automated text analysis systems.

2. Definition

What are Named entities, really? How to define them?

Named Entities in the world: the problem of classification

- **Choice of categories**



- **Definition of the coverage of categories**

Catégorie PERSONNE :

Lionel Jospin	les Démocrates	Bison Futé
les Windsors	les Talibans	le Prince Charmant
la famille Kennedy	Zorro	l'épouse Chirac
les frères Cohen	St Nicolas	…

⟶ unstable categorization

Named Entities in the text: the problem of annotation

- **Combinations of phrases: one or more entities?**

  American and European central banks have decided .....
  Bill and Hillary Clinton
  Utrecht University

- **One phrase: which boundaries ?**

  the democratic presidential candidate Joe Biden, Professor Paolucci
  George W. Bush Jr., La Mecque, Abb Pierre

- **An entity: which lexical units?**

  Jacques Chirac, Mister Chirac, the President Jacques Chirac,
  the French President, the President of the French Republic, 'Chichi'

$\longrightarrow$ very imprecise characterization, diversity of mentions

Named Entities in language: the problem of polysemies

- **Homonymy**

  Orange invited Mr. Holland.

- **Metonymy**

  Leclerc closed its groceries in Rhne-Alpes.

- **"Facets"**

  The candidate Sarkozy, now head of state, has changed his position on the French presence in the international force.

$\longrightarrow$ poly-referentiality

An NLP object difficult to pin down

- **Heterogeneity of achievements**

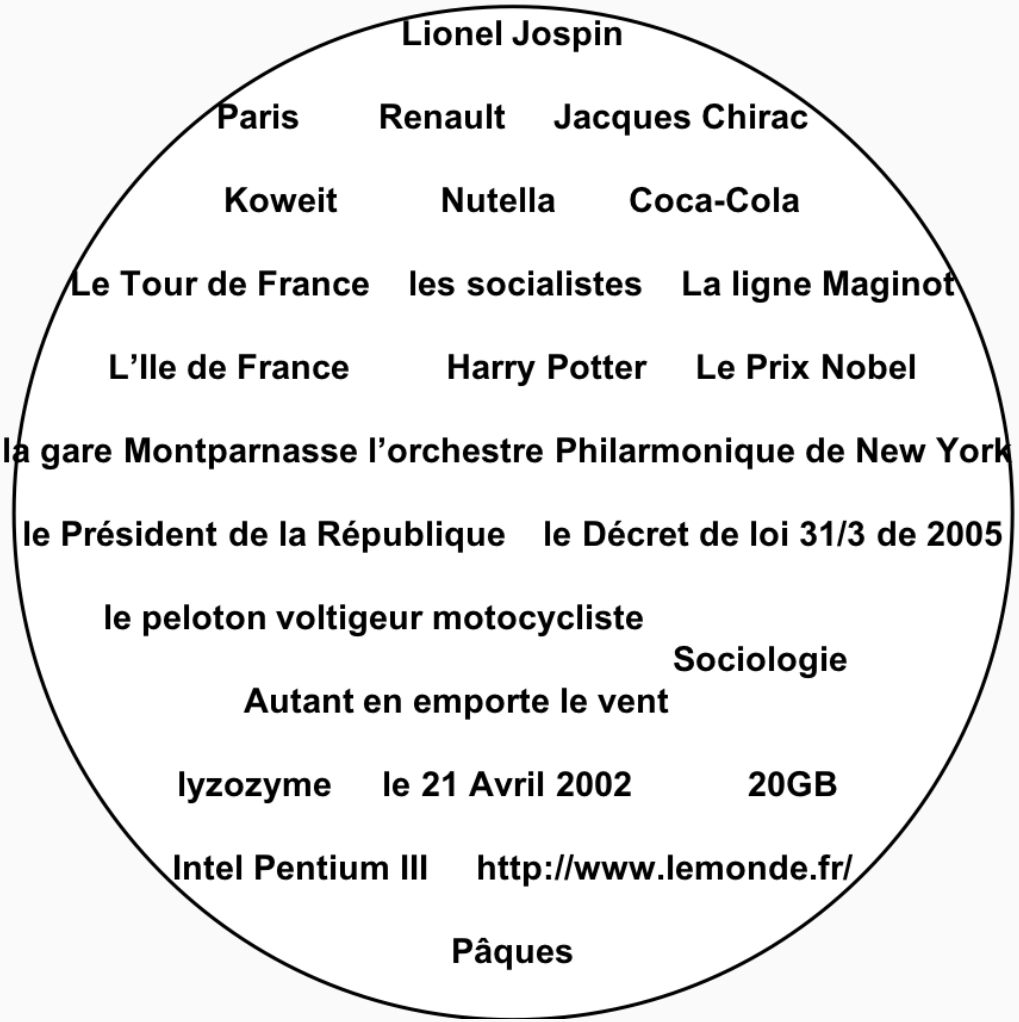
  Named entities are not limited to a categorization, a type of mention, a type of interpretation.

- **Heterogeneity of view points**
  - Definitions in the form of enumerations
  - Various characterizations (considering form or meaning)

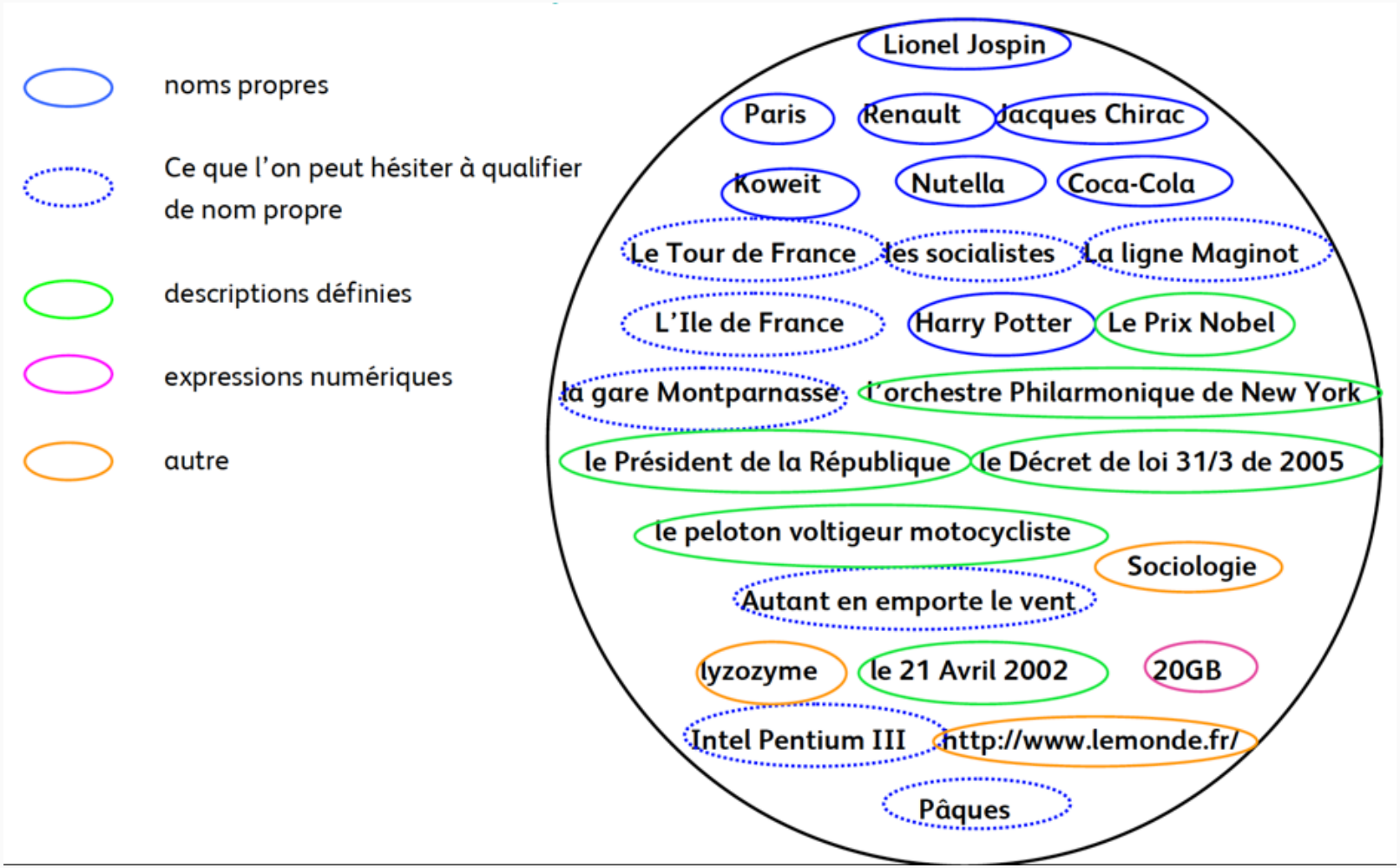
$\longrightarrow$ Question : What are named entities ?

Starting point

Lexical units
considered
as named entities



Starting point



Proposal for a definition

Given an application model and a corpus, we call named entity any linguistic expression that refers to a unique entity of the model autonomously in the corpus.

Questions we asked to arrive at this definition:

- How to define an NLP object ?
- What are proper names ?
- What happens to the notion of meaning and reference in NLP?

Consideration of linguistic aspects

Given an application model and a corpus, we call named entity any linguistic expression that refers to a unique entity of the model autonomously in the corpus.

Referential uniqueness

- **Proper noun refers to individuals**
  - **naming of an individual** (Felix) vs. naming of a class (cat)
  - **uniqueness** : an individual considered as unique within a category of entities
  - **unicity**: an individual considered as a recognizable whole
- **Definite descriptions**
  - presupposition of existence and uniqueness

the president of the Republic, the father of Elisabeth II, the chesnut tree

A denoting phrase in the form "the X" presupposes that there exists one and only one entity that is such as X.

Referential autonomy

How does the reference to a unique entity work?

Proper names

- instructional, denominative meaning → knowledge of a convention
- non contingent naming → rigid designator
- denomination more or less descriptive (*Massif Central*)

Definite descriptions

- descriptive meaning
- proper and improper definite descriptions ()complete/incomplete)

  the president, the President of the French Republic in 2003

Linguistic characterisation

- **The group named entities cannot be reduced to a linguistic category**
  'More than proper names, less than definite descriptions'
- **Specification of a referential behavior**
  Reference to a unique entity and referential autonomy
  Jacques Chirac, the President of the Republic, the blue suit of the president

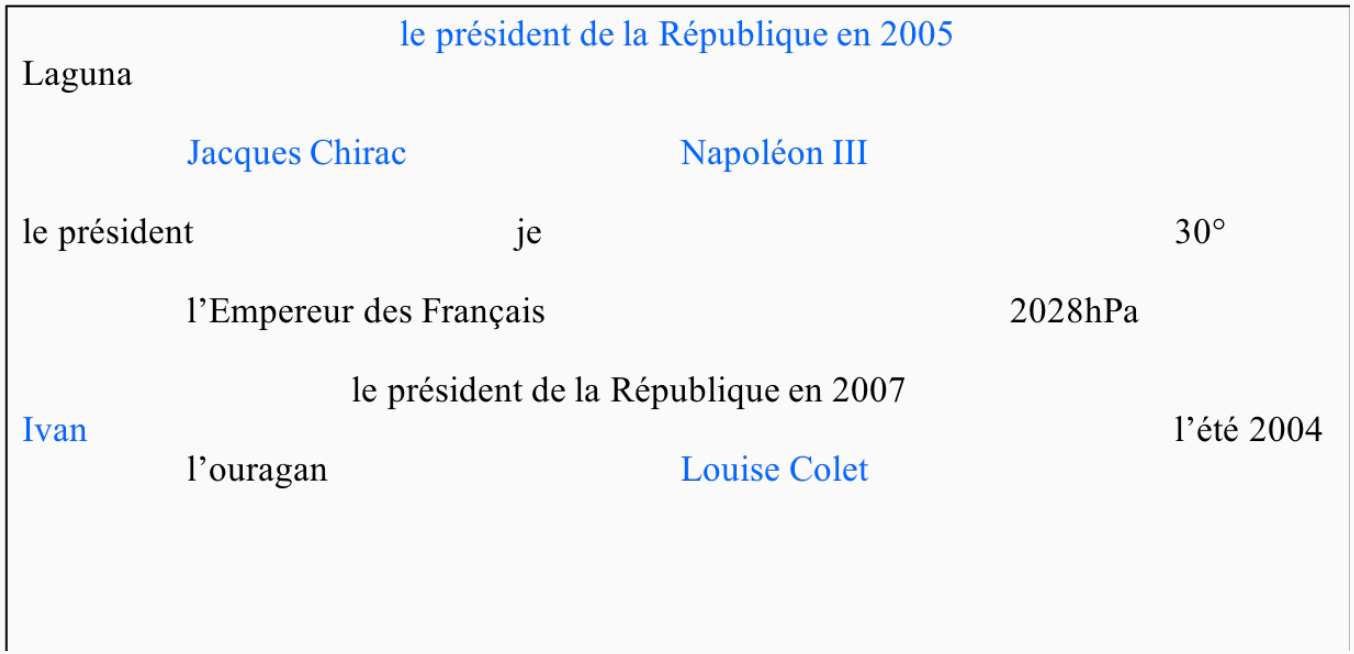
$\longrightarrow$ the linguistic perspective is not sufficient

Consideration of NLP aspects

Given an application model and a corpus, we call named entity any linguistic expression that refers to a unique entity of the model autonomously in the corpus.

Illustration

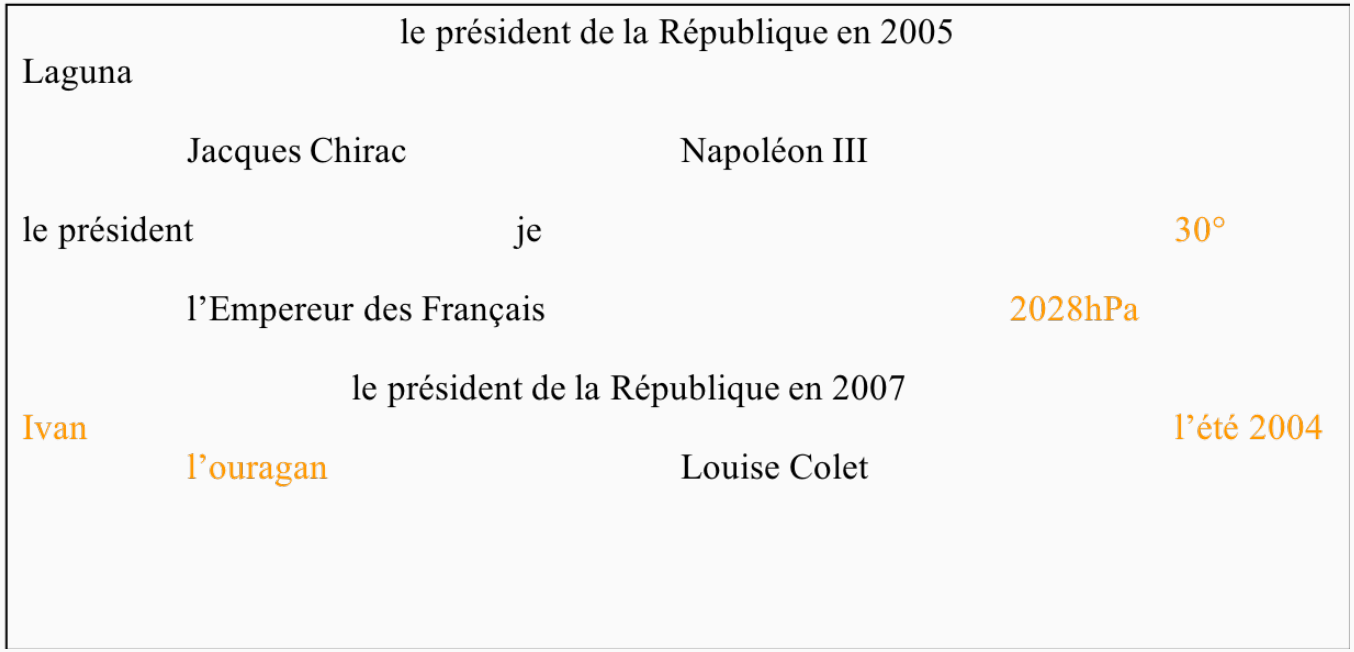
Given an application model and a corpus, we call named entity any linguistic expression that refers to a unique entity of the model autonomously in the corpus.



Application : générique « typique »
Modèle : Personnes, Lieux, Organisations
Corpus : journalistique français de 1998 à 2008

Illustration

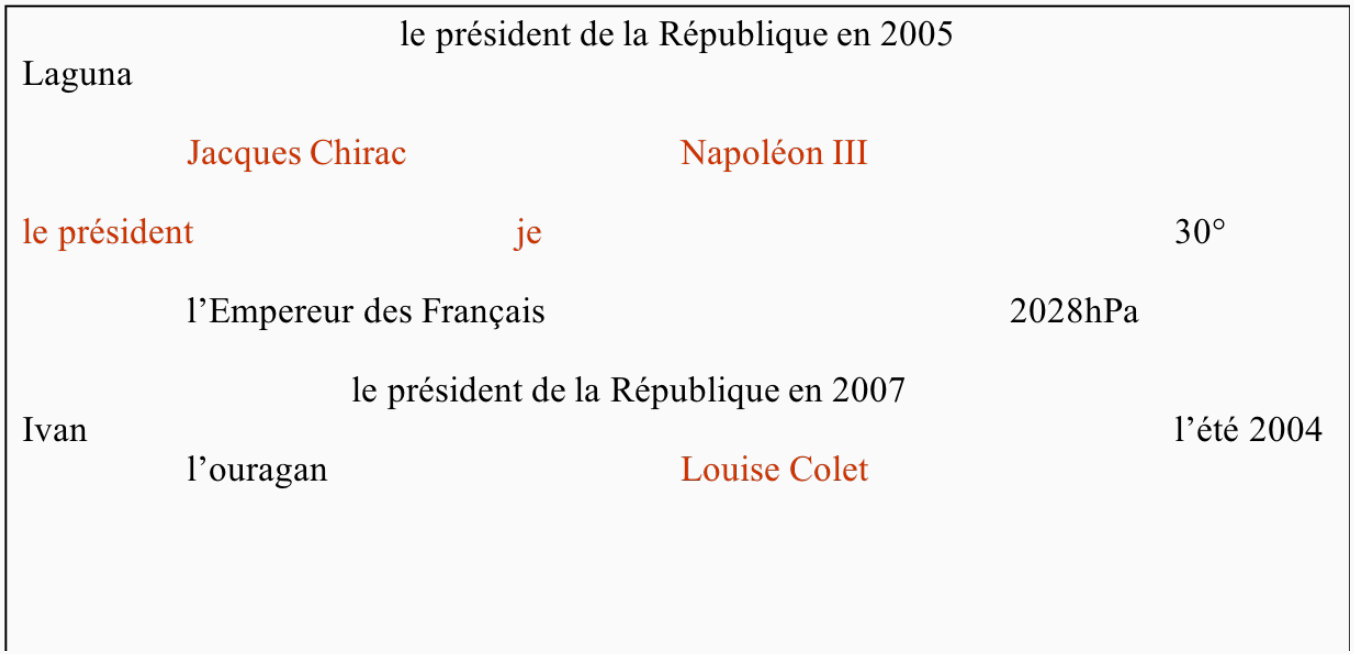
Given an application model and a corpus, we call named entity any linguistic expression that refers to a unique entity of the model autonomously in the corpus.



Application : étude sur le climat
Modèle : températures, mesures atmosphérique, ouragan, dates, périodes, …
Corpus : totalité des observations météorologiques sur une période données

Illustration

Given an application model and a corpus, we call named entity any linguistic expression that refers to a unique entity of the model autonomously in the corpus.



Application : « littéraire »
Modèle : personnes, lieux, événements
Corpus : correspondance de Flaubert

Named entities: an NLP 'creation'

Form linguistics to NLP, specification of a theoretical framework for NEs :

- linguistic perspective: cannot be reduced to a category but can be characterized by a reference behavior
- NLP perspective: exist in relation to a given application model

→ No named entities "per se", only linguistic criteria and a model.

Consequences

- from a general viewpoint : explanation of the heterogeneity and variability of the set 'named entities'
- from a practical viewpoint: decision criteria to annotate
- from a methodological viewpoint: imperative need to explain the model

3. Resources

Resources

What is needed to process named entities?

1. **Typologies**, to define a semantic framework
2. **Annotated corpora**, to serve as a reference (evaluation) and as illustration (training)
3. **Gazetteers and knowledge bases**, to provide background information (training)

3. Resources

3.1 Typologies

Typologies, or how to structure

- A typology (or *tagset*) is a formalized and structured description of semantic categories to consider:
  - which objects of the world should be considered
  - along with a definition of their scope (their realization in texts)
- Typologies differ according to domains and applications
  - different categories
  - different structures, or hierarchies
  - different category definitions

How to define categories?

Approaches:

- top-down: one has a need, defines categories, and that's it.
- bottom-up: categories are deduced from data
- mixte: iteration between the two
- usage of external resources: Wikipedia infoboxes, DBPedia classes (200 classes), more recently Wikidata.

In reality:

- very few explanations about the definition of typologies
- influence of domain, application, funders
- only Sekine [**?**] detailed its methodology to define a typology (200 catgories!).

MUC typologies

- **Proper names** (ENAMEX) : location, persons, organisations,
- **Numerical expression** (NUMEX) : dates and time (expressions absolute expressions), money amount and percentages.

Types	Example	Counter-example
ORG	**DARPA**	our university
PERS	**Harry Schearer**	St. Michael
LOC	**U.S.**	53140 Gatchell Road
MONEY	**19 dollars**	it costs 19
TIME	**8 o'clock**	last night (+ MUC7)
DATE	in **July**	last July (+ MUC7)

Typology ACE

- Recognition and classification of entities, **named or not**, i.e. proper names, nominal phrases, pronouns. $\rightarrow$ detection of all entity mentions.
- Four new categories compared to MUC:
  - **Geo-political Entity** (gpe)
  - **Facility** (fac)
  - **Vehicle** (veh)
  - **Weapon** (wea)
- Introduction of a hierarchy (pers ¿ individuals, groups, undefined) ;
- Distinction between numerical (NUMEX) and temporal (TIMEX) expressions .

N.B: There are several ACE typologies (several iterations)

Typologie ACE

Types	Sub-types
PERS	individual, group, undefined
ORG	governmental, commercial, education, non-governmental, entertainment, media, religious, medical, sciences, sports
GPE	continent, nation, state or province, department or region, cities, special and also pers, loc, org
LOC	addresses, frontiers, astronomic objects, water plans, geographical regions, international regions, others
FAC	airports, factories, facilities
VEH	air, land, water, part of a vehicule
WEA	blunt, explosives, chemical, biologic, guns, bullets, nuclear

Evolution

Several other typologies inspired by MUC and ACE:

- CoNLL: inspiration from MUC, addition of MISC
- HAREM: inspiration from ACE, addition of
  - **Idea** (*abstraccao*): School (*escola*), Field (*disciplina*), Ideology (*ideia*
  - **Object** (*obra*)
  - **Other** (*variado*): close to CoNLL *misc*
  - **Group**, applied to other categories : Title (avec "title group" *grupocargo*), Person (avec "person group" *grupoind*) and Member (*grupomembro*).
- ESTER-2: even more sub-types (e.g. pers.hum, pers.anim, loc.geo, loc.admin, etc) and consideration of nested entities

Refs: [**?**, **?**, **?**]

Typology ESTER-2

Types	Sous-types
pers	pers.hum, pers.anim
fonc	fonc.pol fonc.mil fonc.admi fonc.rel fonc.ari
org	org.pol org.edu org.com org.non-profit org.div org.gsp
loc	loc.geo loc.admi loc.line loc.addr (+3) loc.fac
prod	prod.vehicule prod.award prod.art prod.doc
time	time.date (+ 2 abs et rel) time.hour (+ 2 abs et rel)
amount	amount.phy.age amount.phy.dur amount.phy.temp
	amount.phy.len amount.phy.area amount.phy.vol
	amount.phy.wei amount.phy.spd amount.phy.other
	amount.cur

Nested entities

Besides hierarchization, there is also nested entities:

- one entity may contain another
- *The ¡pers¿ president of ¡org¿ Ford ¡/org¿ ¡/pers¿*

Intensive usage of nested entities in very specilized domains
e.g. GENIA typology in bio-medical [**?**].

Fine-grained characterisation

At the end of the 2000s, the **Quaero program** defined a new typology, used in the ETAPE campaign:

- inspired from ACE for main categories
- decomposition of the typology (and of the task) into two levels:
  1. characterization of types and subtypes (*types*)
  2. characterization of the units making up the NEs (*components*)

$\longrightarrow$ hierarchical typology and compositional entities

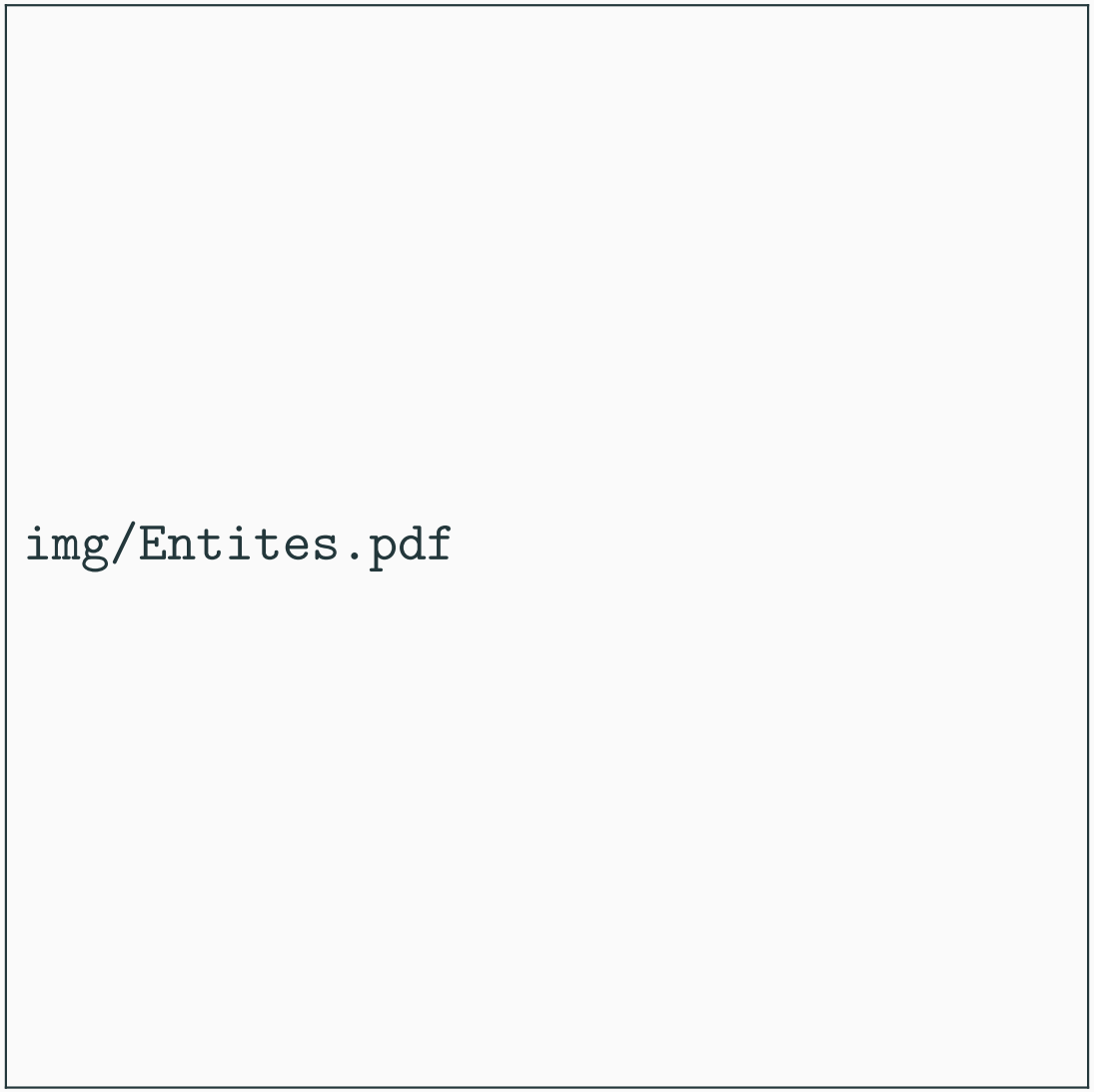
Ref: [**?**, **?**]

Quaero typology

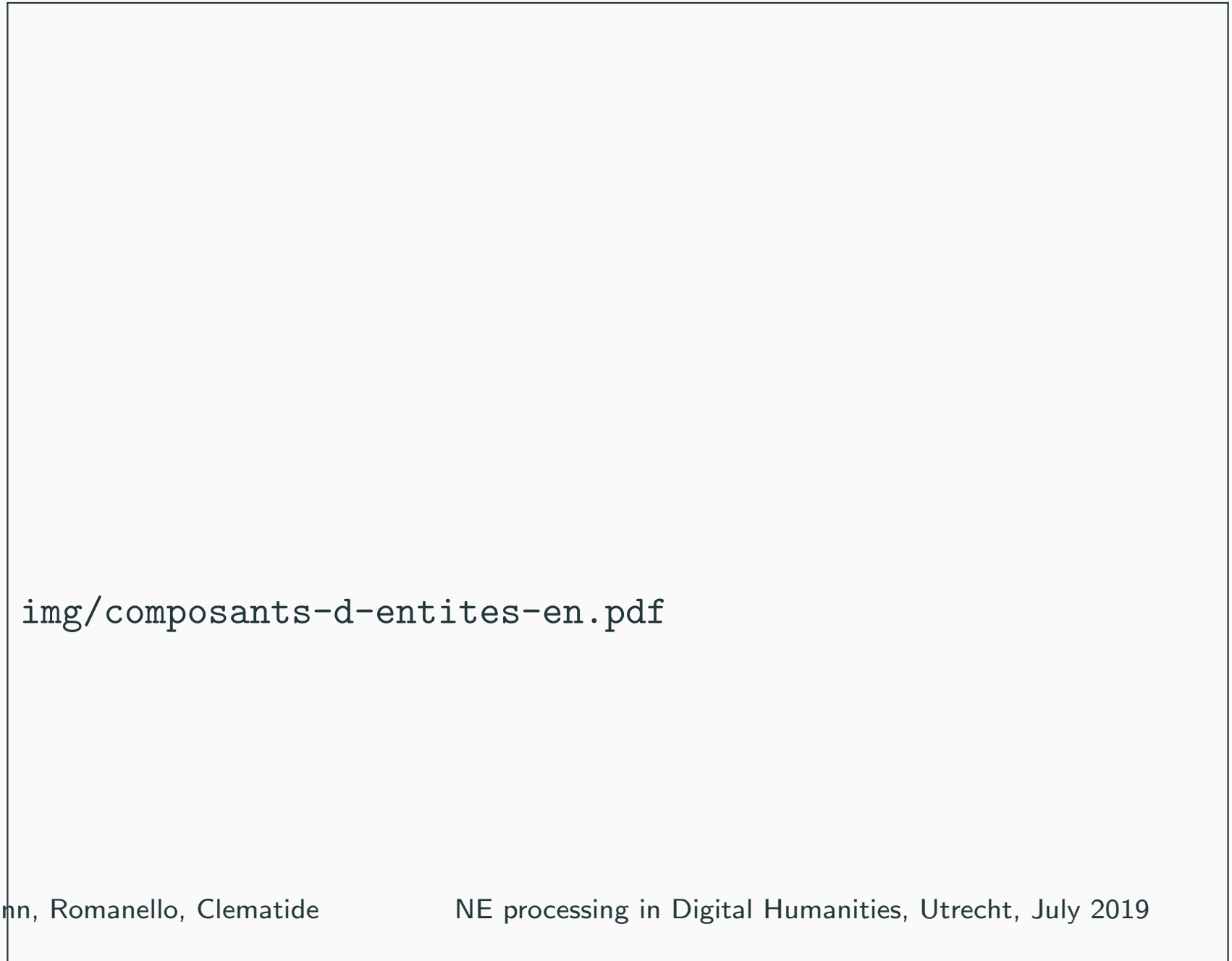
8 main categories

- **Person**: individual person, group of persons;
- **Location**: administrative location, physical location, facilities, oronyms, address;
- **Organization**: administration, service;
- **Time**: absolute and relative date, absolute and relative hour;
- **Amount**;
- **Product**: manufactured object, transportation route, financial products, doctrine, law, software, art, media, award;
- **Function**: individual function, collectivity of functions;

Quaero typology: sub-types



Quaero typology: components



Comparison of typologies

MUC d'aprs le Bureau du recensement des LOC*[Etats-Unis]* , les revenus des mnages ont recul pour la quatrime anne conscutive en DATE*[2011]* .

ACE d'aprs le ORG*[Bureau du recensement des Etats-Unis]* , les revenus des mnages ont recul pour la quatrime anne conscutive en DATE*[2011]* .

ESTER d'aprs le ORG*[Bureau du recensement des LOC[Etats-Unis]]* , les revenus des mnages ont recul pour la quatrime anne conscutive en DATE*[2011]* .

QUA d'aprs le ORG *[name [Bureau du recensement] des LOC [name[Etats-Unis]]]* , les revenus des mnages ont recul pour la quatrime anne conscutive en DATE*[year[2011]]* .

Main issues and points of divergence

Typologies offer different answers to:

- **Management of metonymy**: shall we annotate the contextual or the abolute meaning of an entity ?
  e.g. *France* can refer to the country, the political organization, the sports team.

- **Management of nested entities**

- **Management of coordinated entities**
  Michelle and Barack Obama: one or two entities?

Since 2009 : Text Analysis Conference Knowledge Base Population (TAC-KBP)

Given an entity, one should find its attributes.

E.g. for the type PERS:

- **names** : other names of the person (alias, fake names, stage name, *etc.*) ;
- **functions** and **activities**: jobs, occupations, *etc.* ;
- **dates** (or age): birth, death, different life events, age ;
- **locations**: places related to life events such as birth, death, but also jobs *etc.* ;
- **related persons** : spouse, children, family members, *etc.* ;
- **other information** : alma mater, visited countries, *etc.*

$\longrightarrow$ back to text understanding !

Back to text understanding

- The task remains very complex despite (significant) progress
  - in 2010 the best system did not exceed 0.30 F-measure, in 2014 the highest score was 0.36 [**?**]
- NEs stay at the heart of the process:

This year's slot filling evaluation represented an effort at continuity (. . .) It remains difficult to achieve F-measure higher than 30%. Reaching competitive performance on this task requires a fairly mature NLP system, such as **high-quality name tagging**, *coreference resolution and syntactic analysis.* [**?**]

Back to text understanding

- Resources such as DBpedia do not have sufficient coverage

The TAC-KBP query set contains 3904 entity mentions for 560 distinct entities; entity type was only provided for evaluation. The majority of queries were for organizations (69%). **Most queries were missing from the KB (57%)**. *77% of the distinct GPEs in the queries were present in the KB, but for PERS and ORG these percentages were significantly lower, 19% and 30% respectively* [**?**]

Typologies in DH

TODO

tyologies: take-home message

- **Essential** to the task of NERC
- Strong heritage of **MUC** and **ACE**
- Large **variety** (more than 20 typologies inventoried in 2016)...
- ...but always the **universal triade** (person, organisation, location)
- Trend towards **complexification** (nesting, components, knowledge base population)

Typologies define the framework for action.

They are essential to the creation of *corpora*.

3. Resources

3.2 Annotated corpora

Annotated corpus

A set of textual documents enriched by named entity tagging according to a given typology during an annotation campaign.

Annotation guidelines

Typologies $\longrightarrow$ Annotation guide

- Exemplification of categories
- Rules to allow the annotator to make choices
- Often, parallel definition of the typology and of the annotation guide

Annotation campaign

- Using dedicated softwares (BRATT, GLOZZ, WEBANNO, INCEPTION)
- Importance of measuring the quality and consistency of annotations
- Publication of the corpus with information about: sources, inter-annotator agreement, measures used, typology and annotation guide.
- To do with care: time and resource consuming !

Examples of English corpora

-
-
-

Examples of German corpora

-
-
-

Examples of Italian corpora

-
-
-

Examples of Greek and Latin corpora

-
-
-

Examples of French corpora: QUAERO

1. **Speech transcription corpus**
   - ESTER 2 + other documents
   - Quaero typology
   - 2 sub-corpus: training & test
   - ca 120,000 mentions
   - ELRA-S0349
2. **Historical newspaper corpus**
   - newspapers from 19c (output OCR)
   - Quaero typology
   - 2 sub-corpus: training & test
   - ca 150,000 mentions
   - ELRA-W0073

In Digital Humanities

-
-
-

Overview of existing (non DH) corpora

Inventory of ca. 160 corpus in 2016, with different:

- languages (but prevalance of l'**english**)
- domains (but prevalance of **general**)
- modalities (but prevalance of **written**)
- typologies
- formats
- licenses
- building methodologies

corpus: take-home message

- essential for training and evaluation
- close link with typologies
- expensive to develop
- dominance of Western European languages and of the written modality

Reminder

What is needed to process named entities?

1. **Typologies**, to define a semantic framework
2. **Annotated corpora**, to serve as a reference (evaluation) and as illustration (training)
3. Gazetteers and knowledge bases, to provide background information (training)

3. Resources

3.3 Gazetteers and Knowledge bases

Gazetteers and Knowledge bases

Objective: provide information relating to entities which may be used by automatic systems for the purposes of recognition, classification and disambiguation.

2 types of information:

- **lexical**, on lexical units composing entities
- **encyclopaedic**, on entity referents

A central element for the recognition and classification of NEs up to now, moving to the background with deep learning.

Gazetteers and Knowledge basess

Significant evolution of these resources since the 90s:

basic ‘gazetteers’ $\longrightarrow$ bases encoding of more and more information

Lexica

Encode 2 types of information:

- **entity names or parts of names**, with their associated types
  e.g. *Justin Trudeau*
  → to use in look-up procedures

- **trigger words**, with their associated types
  e.g. *Sir*
  → to use as features to guess names in texts

Building lexica

- High **dependence on application domain**
  e.g. trigger words for general vs. bio-medical domain
- **Challenge 1**: favour **quality** over quantity:
  a small number of entries is enough to recognize a majority of entities
- **Challenge 2**: comply with the **rapid evolution** of NEs that are an open class

PROLEX

- Multilingual lexical base for named entities, which contains
  - entities ('pivots')
  - + their surface forms ('prolexmes')
  - + their types
  - + their relations (e.g. synonymy, meronymy)
  - v2.2: French (100k units), English, Polish
- Developed by University Franois Rabelais of Tours, France

PROLEX

CNRTL — Centre National de Ressources Textuelles et Lexicales

Ortolang — Outils et Ressources pour un Traitement Optimisé de la LANGue

cnrs atilf

Accueil | Portail lexical | Corpus | Lexiques | Dictionnaires | Métalexicographie | Outils | Contact

Prolex

Le projet *Prolex*, piloté par le Laboratoire d'informatique (LI) de l'université François-Rabelais de Tours, a pour but de fournir, à la communauté du traitement automatique des langues (Tal), des connaissances sur les noms propres, qui constituent, à eux seuls, 10% des textes journalistiques. Ceci par la création d'une plate-forme technologique comprenant un dictionnaire électronique relationnel multilingue de noms propres (*Prolexbase*), des systèmes d'identification des noms propres et de leurs dérivés, des grammaires locales, etc.

La ressource Prolexbase est un projet Tal du LI, en collaboration avec :

- le laboratoire ligérien de linguistique ;
- l'université de Belgrade;
- l'académie des sciences de Varsovie.

Ce projet a reçu le soutien :

- de l'action Technolangue du Ministère de l'Industrie (2003-2005) ;
- du programme d'action intégré Egide Pavle-Savic du Ministère des Affaires étrangères (2004-2005) ;
- du projet Feder Région Centre Entités nommées et nommables (2009-2010) ;
- du projet ERDF Nekst (2009-2014) ;
- du projet européen (CIP ICT-PSP) Cesar (2011-2013).

Prolexbase

La modélisation du domaine des noms propres définie dans le projet Prolex repose sur deux concepts centraux : le *pivot* et le *prolexème*. Le pivot ne représente pas le référent, mais un point de vue sur ce référent. Il possède dans chaque langue un concept spécifique, le prolexème, qui est une famille structurée de lexèmes. Autour d'eux, sont définis d'autres concepts et des relations (synonymie, méronymie, accessibilité, éponymie, etc.). Chaque pivot est en relation d'hyperonymie avec un type et une existence au sein de deux typologies.

Il n'est pas évident de définir la notion de nom propre. La plupart des définitions insistent sur le caractère unique de son référent et sur une sémantique et une syntaxe qui lui est propre. Nous avons choisi d'adopter le point de vue de (Jonasson, 1994) qui propose une définition plus large incluant ce qu'elle appelle les noms propres purs (noms de personne et noms de lieu) et les noms propres descriptifs qui résultent souvent de la composition d'un nom propre avec une expansion (Tour Eiffel, musée Rodin, etc.). Un nom propre descriptif peut être considéré comme une expression définie figée ou en cours de figement (Jardin des Plantes, Médecins sans frontières, etc.). Cette définition est assez proche de celle utilisée dans le domaine du Tal depuis la conférence MUC6.

Origine de la ressource	LI (Université François-Rabelais de Tours)
Nature des données	Lexique relationnel multilingue de noms propres
Soutiens institutionnels	Action Technolangue du Ministère de l'Industrie Programme d'action intégré Egide Pavle-Savic du Ministère des Affaires étrangères Projet Feder Région Centre

`http://www.cnrtl.fr/lexiques/prolex/`

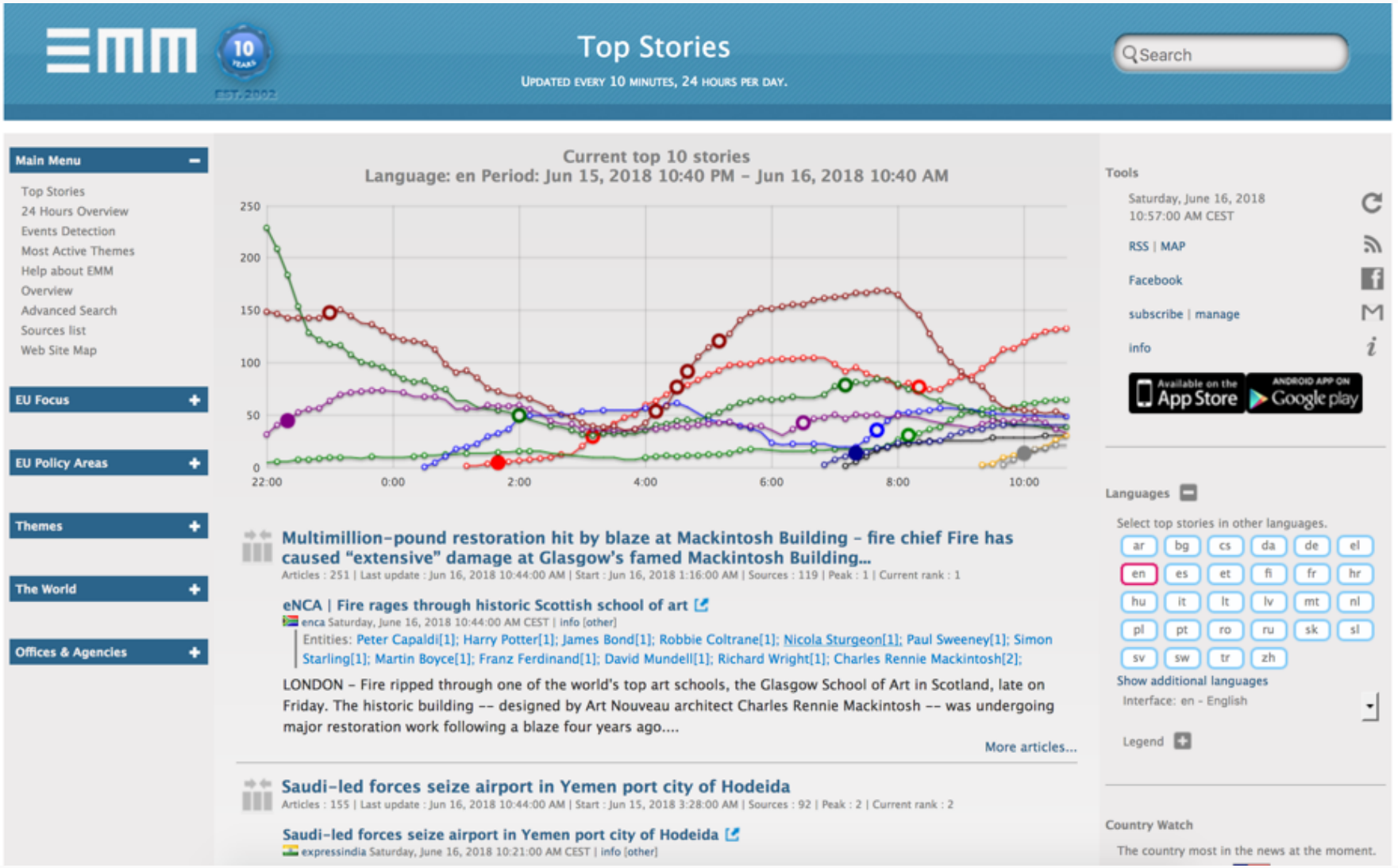
GEONAMES

- toponyms
- 7 millions entities and 10 millions lexical entries (variants)
- properties: coordinates, population, postal code, etc.
- assignment of an URI to each entity
- 9 main types (divided into 645 sub-types)

JRCNames

- a 'by-product' of a media monitoring system:
  7000 sources, 300k articles per day, 70 languages, among which 21 with NE processing
- ca. 340,000 unique entities (PERS et ORG)
- 1,7 million name variants (lexicalisations) in 170 languages
- 32 millions relations (cross-lingual)
- up to 400 variants for one entity

JRCNames



http://emm.newsbrief.eu

JRCNames

Main Menu

- Top Stories
- 24 Hours Overview
- Events Detection
- Most Active Themes
- Help about EMM
- Overview
- Advanced Search
- Sources list
- Web Site Map

EU Focus +

EU Policy Areas +

Themes +

The World +

Offices & Agencies +

Nicola Sturgeon

Last updated on 2018-02-21T08:07+0100.

About this image
License: UNKNOWN
Author: The Scottish Government

Extracted quotes from

Nicola Sturgeon said : "not listened to, who is responsible and how are we going to ensure individuals are accountable?"
thecourier Thursday, June 14, 2018 6:35:00 PM CEST

Nicola Sturgeon said : "Yesterday morning I was spending my time in two primary schools, as well as a secondary school and an early years centre. And I was talking to a range of primary school children including some five-year-olds. "I didn't meet any of them in tears, it didn't see any of them that looked crushed. What I saw were confident, bright enthusiastic young people – some of those were showing me computer coding and some were speaking Mandarin, that is how confident they were"
bbc Thursday, June 14, 2018 4:32:00 PM CEST

Nicola Sturgeon said : "As local MSP for the Gorbals, I'm in close contact with NewGorbalsHA who are on site. "
dailystar Thursday, June 14, 2018 1:53:00 PM CEST

Nicola Sturgeon said : "As local MSP for the Gorbals, I'm in close contact with NewGorbalsHA who are on site. "
dailystar Thursday, June 14, 2018 12:21:00 PM CEST

Key Titles and Phrases (Last 30) | Names (Top 30)

Key Titles and Phrases	Count	Lang	Last Seen
minister	47.38%	EN	15/06/2018
leader	9.97%	EN	15/06/2018
première ministre écossaise	4.26%	FR	15/06/2018
ministre écossaise	3.87%	FR	15/06/2018
first minister of scotland	1.72%	EN	14/06/2018
minister of scotland	1.06%	EN	14/06/2018

Related entities (Top 30) | Associated entities (Top 30)

Type	Entity Name	Count
	EU	7.23%
	Glasgow School	5.40%
	Charles Rennie Mackintosh	5.30%
	Ian Blackford	4.18%
	Theresa May	4.18%
	Paul Sweeney	3.97%

1 2 3 4 5 6 7 8 9 10 > »10

Articles published more than 12 hours ago

Tools

Saturday, June 16, 2018
10:58:00 AM CEST

Facebook

manage

Available on the App Store | ANDROID APP ON Google play

Languages

Select your languages

am ar az be bg bs
ca cs da de el en
eo es et fa fi fr
ga ha he hi hr hu
hy id is it ja ka
km ko ku ky lb lo
lt lv mk ml mt nl
no pap pl ps pt ro
ru rw si sk sl sq
sr sv sw ta th tr
uk ur vi zh
all

Interface:
en - English

Legend +

Explore Relations

Extracted quotes about

Adam Tomkins said (about Nicola Sturgeon) : "This is a remarkable report which exposes Nicola Sturgeon's secret Scotland. "People will see this report and

JRCNames

- published in format .txt in 2011 [add ref]
- RDF in 2016 [add ref]

Lexicons in DH?

Conclusion on lexica/gazetteers

- Not all are listed or published
- Initially: describe possible linguistic realizations of entities
- Evolution: enrichment with further information, e.g. date of birth of a person, population of a city, etc.
- 3 enrichment perspectives:
  - larger coverage
  - multilinguism
  - encyclopaedic information

$\longrightarrow$ more complex and larger data structures

Knowledge bases (quick overview)

- **Wikipedia** (initiated in 2001)
  - useful for extracting and integrating NE lexicons
  - semi-automatic constitution of annotated corpora
  - acquisition of relations between entities
- **DBpedia** (RDF equivalent of Wikipedia)
- **YAGO** (Wikipedia, WordNet, with spatial and temporal info)
- **BabelNet**
- **Wikidata**
- **OpenCyc** (free part of Cyc), information about 'common sense'

lexicas and knowledge bases: take-home message

- third pillar in terms of resources for NEs
- lexical and semantic information
- difficult to acquire, represent, store and use until mid-2000
- today: information explosion, mainly for the general domain

4. Recognition et classification

NE processing - task reminder

1. recognition: detecting, spotting named entities in textual streams (one delimitates NEs 'boundaries' in texts)

2. classification: categorizing detected segments according to pre-defined semantic categories (one assigns a type)

3. disambiguation/linking: linking entity mentions to a unique reference (one determines the reference)

4. relation extraction: discovering relations between NEs (e.g. *father-of, born-in, alma mater*)

Objective

Build systems that perform these tasks automatically.

Requirements:

- quality: do not do too many mistakes
- exhaustivity: do not miss too many entities
- robustness: do not fail in front of non-canonical cases

In practice:

- difficult to meet these 3 requirements simultaneously
- looking for the alertbest compromise according to resources and application.

Remarks

Starting point: **text**, represented as a linear structure, *i.e.* a sequence of words that can be split into tokens

We wish to detect if the evidence/clues presents in tokens indicate:

- the presence of a named entity
- the presence of a given category
- the reference to a certain entity

Warning: no systematic mapping between a set of properties and a named entity category.

4. Recognition et classification

4.1 Clues

Text representation

Linear text representation as sequence of words features 2 levels of granularity:

- **characters**, which compose words
- **words**, which compose sequences (text)

Clues can appear at different levels:

- characters: **morphological clues**
- words: **lexical clues**
- word sequence: **contextual clues**

Morphological clues: summary

Namely:

- characters which compose words
- existence of different classes of characters
- existence of regularities in the arrangement of characters, some of which are useful for ENs

Interactive session

According to you, which morphological indices can help recognize named entities?

Capitalization

- widely used in Western character sets to mark a proper noun
- very easy to test to detect NEs

BUT

- capitalization also used to start sentences, or in acronyms
- used for common nouns in German
- does not help for classification
- notion of capitalization not widely used in Orient (does not exist in Chinese, Hindi, Arabic, etc.)

Socio-cultural regularities

- the suffix *-ville* or the prefix *Saint-* in French
- regular suffixes for person names
  - in Russian, the suffix *-vitch*
  - in Swedish, the suffix *-sson*
  - in Icelandic , *-dttir*
  - in North Africa, prefixes *Ben-* or *At-*
  - in Japanese, the suffix *-san*
- tokens originating from conventions or standards :
  Inc. in English, *S.A.* in French, *GmbH* in German

Presence of numerical characters or numbers

- dates, amounts or measures typically contain numbers (written in numbers or letters)
- numbers can be written differently (*10 000*, *quatre-vingt-treize*, *cent dix-huit*) and/or with specific characters (*10,38*, *24/03*).
- mix of numeric and alphabetic characters (*100km*, *10h30*, etc.).
- abbreviations and acronyms (*A380*, *ISO-9000*, *Canon EOS 70D*).

Morphological clues

Allow the detection of entities characterized by regularities, for a number of languages

But it is not sufficient:

- only cover standardized NE forms
- allow to *detect*, but not to *classify*

Lexical clues

Principle: confronting texts with lists of entities or entity components.

- powerful and accurate if lexical entries are controlled
- lexicons often organized according to NE types or degree of ambiguity
  e.g. *Hollande* vs. *Obama*
- At stake: finding the right compromise between quantity and efficiency

Lexical clues

In practice: algorithms return the list of textual segments corresponding to lexicon entries.

Warning: a look-up procedure retrieves several lexicon entries for ambiguous tokens.

Today, Franois Hollande met Obama in Washington.

- the Person *Franois Hollande*
- the Location *Hollande*
- the Person *Obama*
- the Person or Location *Washington*.

Lexical clues

Challenge: NE are an open class, it is not possible to be exhaustive

- new names are continually being created [**?**, **?**]
- some parts of defined descriptions are substitutable

$\rightarrow$ keeping an NE lexicon up to date with all of their exact forms and variations is a costly and complex task.

Lexical clues

Often, **joint consideration** of morphological and lexical indices:

- **Persons** : first word capitalized, second is a proper name
- **Dates** : first token is a number, second is part of the list of month names (*5 juillet 2012*)
- **Locations** : preceded by *in* ou *to*, and followed by a river name (*Montlouis sur Loire*)
- etc.

In Digital Humanities

- validity of lexical clues
- additional difficulties: historical spellings, historical name variants and trigger words (e.g. occupation names)
- ...

Contextual clues

In some cases, words that make up named entities are not sufficient: Morphological and lexical indices may be absent or ambiguous.

$\rightarrow$ need additional clues nearby:

- **local context**: words that precede or follow the entity.
- **global context**: sentence, close sentences, paragraph, document.

Importance of contextual clues

1. *He saw Holland on television.*
2. *His trip to Holland went well.*
3. *He bought a Renault Clio.*
4. *The Clio muse sings the past of humans and cities.*
5. *I'm reading about Washington for my work.*

Since entity spellings are identical, only the context can help classify.
Easy and intuitive for the human, more complicated for a machine.

Contextual clues

- Processing applied to the text, not to words in isolation;
  $\rightarrow$ higher computational cost;
- Often need to rely on preliminary analyzes: syntax, correference, thematic of the document;
- More economic: select, *a priori* or *a posteriori* most discriminating contextual clues and their combinations;
- Analysis of context important when EN typology is fine-grained.

clues: take-home message

- Morphological, lexical or contextual
- Possibility of composite indices
- these are the 'ingredients' of NE processing systems

4. Recognition et classification

4.2 Approches symboliques

Techniques base d'automates

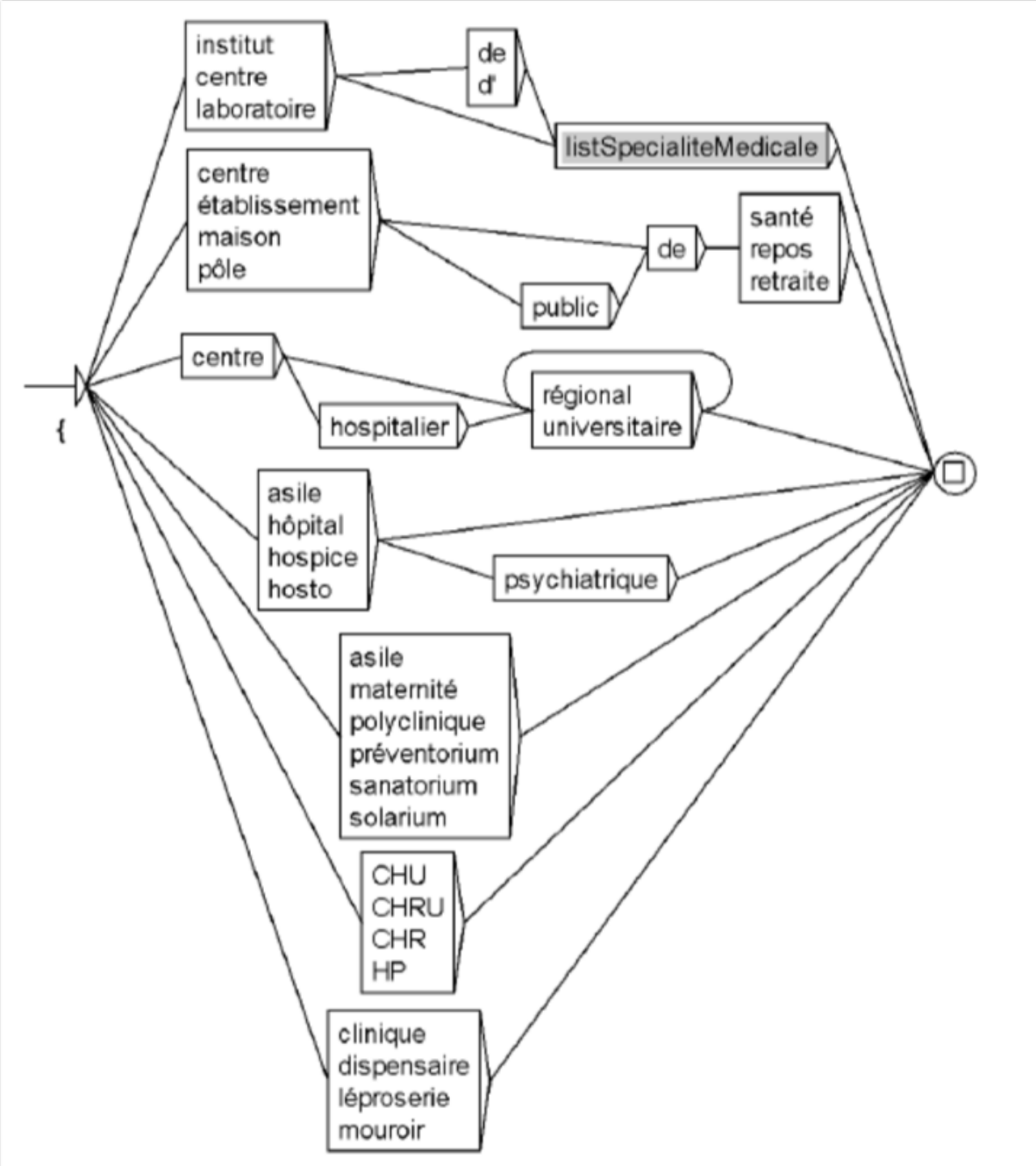
- **Objectif**: insertion de balises dans les textes indiquant o se trouvent les ENs
- **Principe**: conception de *rgles* formant un *grammaire locale*
- **Ralisation**: utilisation de transducteurs
  chaque rgle est associe un diagramme d'tats o les nuds reprsentent les tats de l'automate et les artes les transitions
- De nombreuses **botes outils**:
  - GATE
  - LingPipe
  - NooJ
  - OpenNLP
  - OpenCalais
  - Unitex

Automates

- des lments sont indiqus au sein des nuds
- les noeuds sont agencs de manire reconnatre des expressions linguistiques
- les transitions sont ralises par prsence d'indices (morphologiques, lexicaux, internes ou externes)
- plusieurs transitions sont ralises par juxtaposition de nuds
- l'automate ne reconnat une expression linguistique que s'il existe un chemin depuis le nud initial (gauche) jusqu'au nud final (droite).

Objectif : contraindre correctement l'automate, afin qu'il reconnaisse toutes les expressions linguistiques souhaites, et aucune autre.

Automates


institut
centre
laboratoire
de
d'
listSpecialiteMedicale
centre
établissement
maison
pôle
de
santé
repos
retraite
public
centre
hospitalier
régional
universitaire
asile
hôpital
hospice
hosto
psychiatrique
asile
maternité
polyclinique
préventorium
sanatorium
solarium
CHU
CHRU
CHR
HP
clinique
dispensaire
léproserie
mouroir

Automates

Possibilit d'avoir des prtraitements:
segmentation en mots, en phrases, tiquetage morphosyntaxique.

$\rightarrow$ indices supplmentaires fort utiles,
mais qui impactent les performances si bruits.

Basculement vers les approches statistiques

Au dbut des annes 2000, grce la mise disposition de jeux de donnes volumineux.

Mais les approches symboliques sont toujours prsentes:

- combines avec des mthodes statistiques
- prdominent pour les langues ou les typologies sans corpus de donnes suffisants
- gardent l'atout pour le contrle et de l'ingnierie: plus comprhensibles, modulables, possibilits de rglages fins.
- majoritaires dans le milieu industriel.

4. Recognition et classification

4.3 Modles guids par les donnes et apprentissage

Le paradigme de l'apprentissage automatique

Objectif: dterminer les paramtres d'un modle partir de donnes, d'o le terme *apprentissage*

Ces paramtres et ce modle sont ensuite utiliss pour prendre les dcisions les plus probables (ou vraisemblables) sur de nouvelles donnes traiter.

Il s'agit, simultanment, de spcifier le modle et de gnraliser les donnes.

Le paradigme de l'apprentissage automatique

A partir des annes 1960, mergent les modles connexionnistes (perceptron, rseaux de neurones), qui mettent en relation des proprits sur les objets modliss.

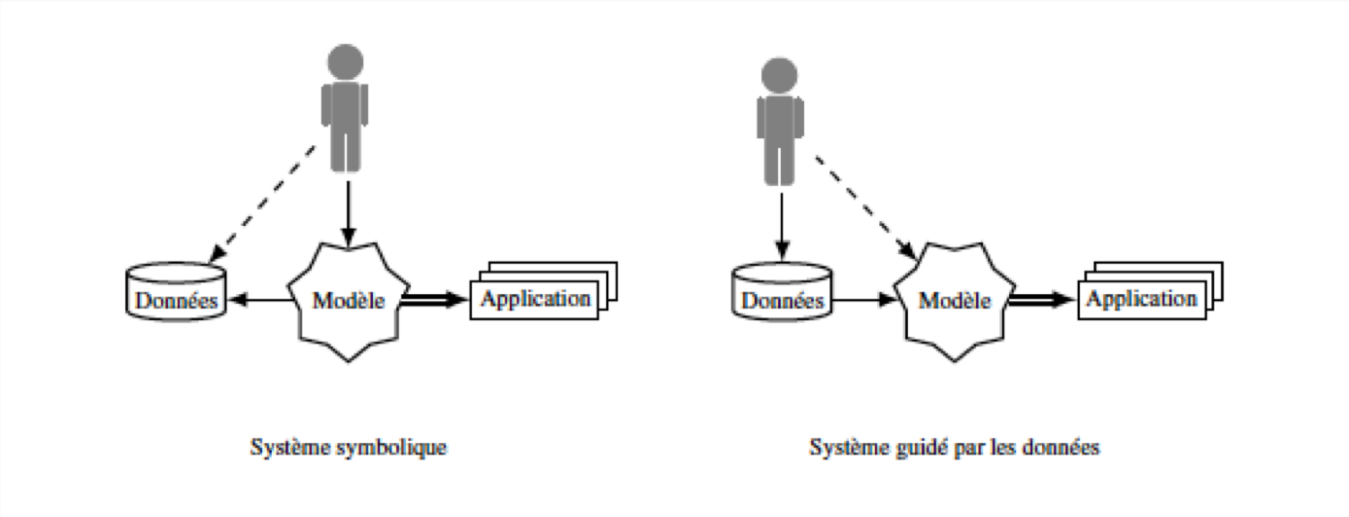
Puis les modles markoviens (modles de Markov tats cachs), qui simulent des processus stochastiques.

$\rightarrow$ remise en cause du principe dterministe des automates et la manire dont sont labors les systmes.

Le paradigme de l'apprentissage automatique

Systmes symboliques: le concepteur du systme interagit majoritairement avec le modle (l'automate), et n'utilise les donnes que pour visualiser ou d'valuaer.

Systmes guids par les donnes: le concepteur agit sur les donnes, la structure du modle est prdfinie et rigide et les paramtres ajusts automatiquement partir des donnes.



Le rle des donnes

- format
- quantit
- qualit

$\longrightarrow$ modles plus ou moins prcis, couvrants et robustes

Le codage des donnes

Formats tabulaires. Aprs une segmentation en mots, une tiquette est associe chaque mot:

- catgorie d'entit nomme (`PERS`, `ORG`, etc.)
- une catgorie spciale, `O` (comme *Outside*)

Problme: entits polylexicales vs entits nommes contigus de mme type. e.g. *Paris Brest*

Le codage des donnes

Solution: catgories clates

- Format **BIO** ($2N + 1$ classes diffrentes):
  - mot qui commence une nouvelle entit (`PERS-B` pour *Begin*)
  - mot qui prolonge une entit polylexicale (`PERS-I` pour *Inside*)
- Format **BILOU** ($4N + 1$):
  distingue aussi les entits composes d'un seul mot ou les derniers mots des entits: *Begin*, *Inside*, *Last*, *Outside*, *Unique*
  $\rightarrow$ de meilleures perfomances si donnes annotes suffisantes

Exemple

En 2008 Hollande prend un vol Rio de Janeiro Los angeles

Balises	`En <DATE> 2008 </DATE> <PERS> Hollande </PERS> prend un vol <LOC> Rio de Janeiro </LOC> <LOC> Los angeles </LOC>`
BIO	`En 2008/DATE-B Hollande/PERS-B prend/O un/O vol/O Rio/LOC-B de/LOC-I Janeiro/LOC-I Los/LOC-B angeles/LOC-I`
BILOU	`En 2008/DATE-U Hollande/PERS-U prend/O un/O vol/O Rio/LOC-B de/LOC-I Janeiro/LOC-L Los/LOC-B angeles/LOC-L`

Modles par classes majoritaires

Dterminer la classe d'un mot partir de la classe qui lui est majoritairement associe dans le corpus d'apprentissage.

Formulation l'aide de probabilits:

- frquence du mot $F(m)$
- frquence d'une tiquette $F(e)$
- frquence de la prsence jointe du mot et de l'tiquette $F(m, e)$

Modles par classes majoritaires

La formule de Bayes et l'estimation statistique permettent de calculer la probabilit d'une tiquette sachant le mot :

$$P(E_i = e|M_i = m) = \frac{P(M_i = m, E_i = e)}{P(M_i = m)} = \frac{F(e, m)}{F(m)}$$

Probabilit d'une tiquettes pour un mot donne =
ratio entre la frquence dans le corpus annot du mot avec une tiquette et la frquence dans ce mme corpus du mot (quelque soit l'tiquette)

Modles par classes majoritaires

Pour une squence de mots et d'tiquettes (hypothse d'indpendance entre les mots):

$$P(E_1, E_2 \dots E_n | M_1, M_2 \dots M_n) = \prod_{i=1}^{n} P(E_i | M_i) = \prod_{i=1}^{n} \frac{F(e, m)}{F(m)}$$

Puis: slectionner la suite d'tiquettes qui, en fonction des mots de l'nonc, maximise cette probabilit.

Complexit restreinte ici: choix de l'tiquette la plus probable pour chaque mot.

Modles par classes majoritaires

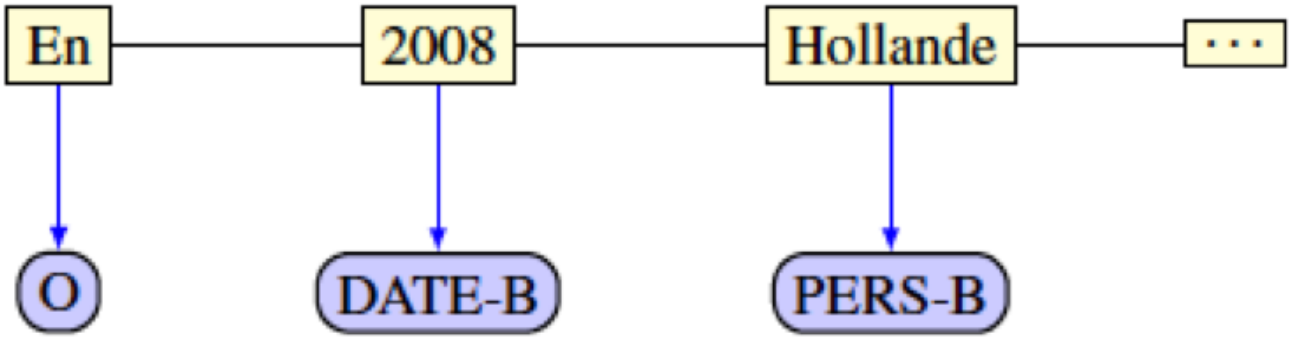


Figure 4.3. *Modèle par classes majoritaires*

(l'orientation des flches indique quelles dpendances sont prises en compte par le modle)

Modles dcisions contextuelles (HMM)

Objectif: tenir compte de la vraisemblance d'tiquettes contigus

Franois Hollande

- *Hollande*: Lieu ou Personne ?
- *Franois*, annot comme Personne, peut conditionner l'annotation du mot *Hollande*

Modles dcisions contextuelles (HMM)

Option: modles gnratifs comme les modles de Markov tats cachs.

Calcul des probabilits invers : dterminer, pour une suite d'tiquettes, la probabilit qu'elle gnre un texte donn.

$$P(M_1, M_2 \dots M_n | E_1, E_2 \dots E_n) = \prod_{i=1}^{n} P(M_i|E_i) * P(E_i|E_{i-1})$$

Soit le produit des probabilits de gnration $P(M_i|E_i)$ et de transition $P(E_i|E_{i-1})$.

Modles dcisions contextuelles (HMM)

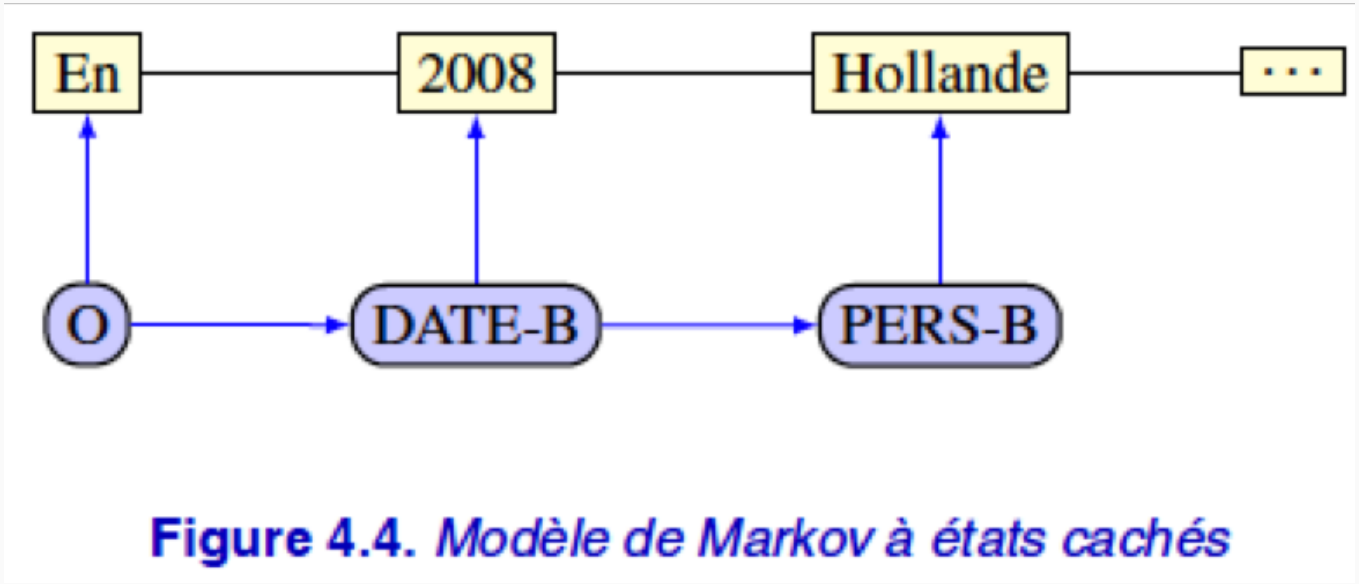


Figure 4.4. *Modèle de Markov à états cachés*

Dcisions non indpendantes : la solution la plus vraisemblable est choisie en fonction des tiquettes pralablement choisies.

Modles utilisant des indices multiples (softmax, MaxEnt)

Objectif: considrer plus d'indices que les mots, i.e. prendre en compte la morphologie, les indices lexicaux, le contexte, etc.

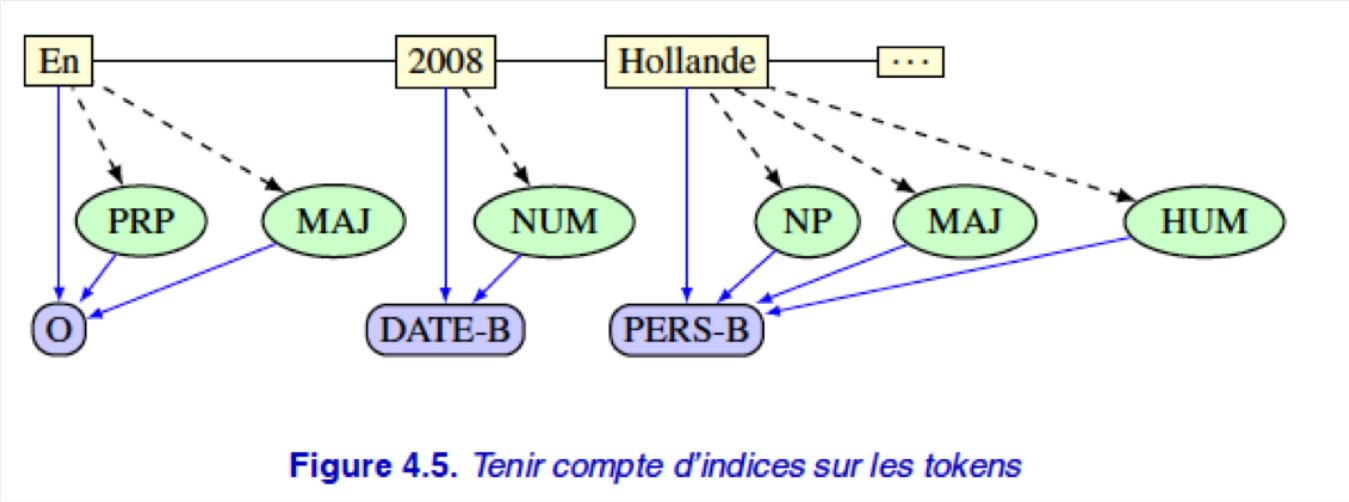


Figure 4.5. *Tenir compte d'indices sur les tokens*

Modles utilisant des indices multiples (softmax, MaxEnt)

Utilisation d'une fonction ddie G, qui utilise des satistiques sur les indices $F_{i1} \ldots F_{ik}$ dans un calcul, ainsi qu' une normalisation parmi les T types d'entits nommes possibles.

$$P(E_i = e | M_i = m, F_{i1} = f_1 \ldots F_{ik} = f_k) = \frac{G(e, m, f_1 \ldots f_k)}{\sum_{t \in T} G(t, m, f_1 \ldots f_k)}$$

Il existe diffrentes manires de dfinir la fonction G.

Champs markoviens conditionnels (CRF)

Les CRF (*Conditional Random Fields* ou champs markoviens conditionnels) combinent les deux aspects prcdents :

- tenir compte du contexte pour prendre des dcisions
  (une dcision sur un mot influence la dcision pour le mot suivant)
- tenir compte de multiples indices
  (analyses en prtraitements)

Modle qui obtient de trs bonnes performances pour la reconnaissance d'EN.

Champs markoviens conditionnels (CRF)

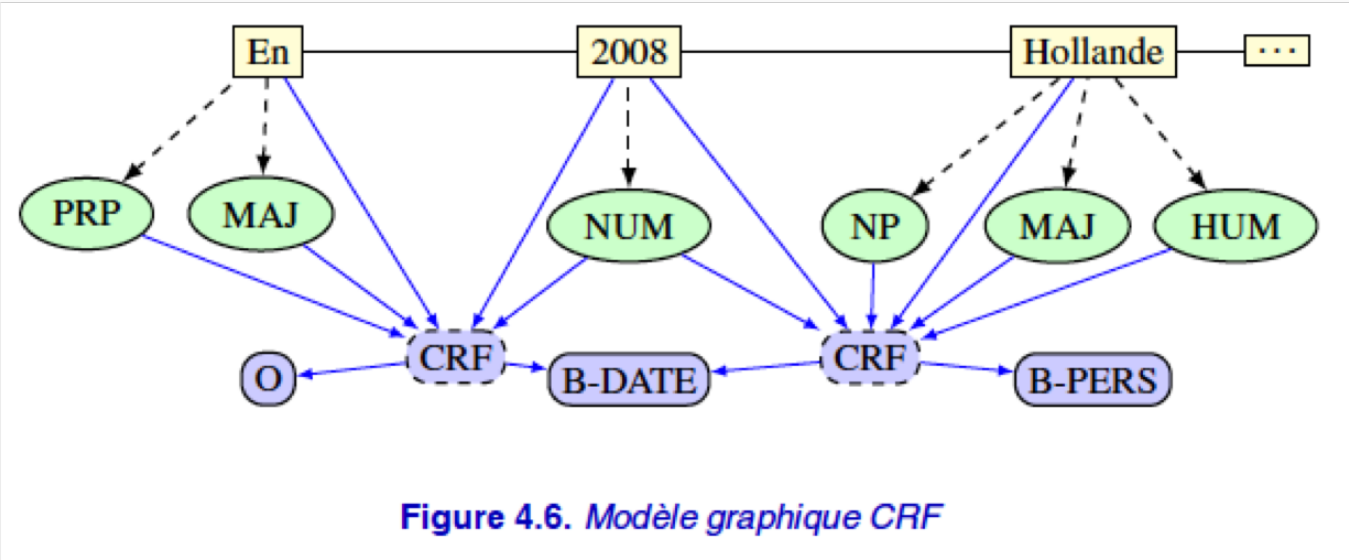


Figure 4.6. *Modèle graphique CRF*

$$G(e, m, f_1 \dots f_k) = exp\left(\sum_{p=1}^{k} \alpha_{ep} * f_p\right)$$

Rseaux de neurones profonds

rcen: botes outils d'apprentissage

- Illinois NETagger
  `http://cogcomp.cs.illinois.edu/page/software/_view/NETagger`
- Stanford NE tagger
  `http://nlp.stanford.edu/software/CRF-NER.shtml`
- ...

Reconnaissance et classification d'en: à retenir

- possibilit d'utiliser de nombreux indices
- via des mthodes diverses qui peuvent tre combines.
- si plus d'indices, alors complexit grandissante et besoin de plus de donnes annotes
- importance de la slection d'indices

(10min)

5. Linking (quickly)

O en sommes-nous

- nous savons reconnaitre et catgoriser des segments textuels: des *mentions* d'entis qui font rfrence un objet du monde.
- ce qu'il reste faire: tablie le lien entre les mentions et les objets auxquels elles rfrent

$\longrightarrow$ objectif: dsambigusation, rsolution, liaison

Des mentions aux rfrents

- **Catgoriser n'est pas dsambiguser**:
  G. Bush et *F. Mitterrand* sont des PERSON
  Mais lequel des 2 rfre au *43me prsident des tats-Unis*?

- **Le problme des homonymes**:
  F. Mitterrand est une PERSON
  Mais *Franois Mitterrand* ou *Frdric Mitterrand* ?

  Bush est une PERSON
  Mais *G. W. Bush* ou *G. Bush* ?

- **Le problmes des variantes**:
  Jean-Claudem Junckerem, *Juncker*, *Jean-Cluade Juncker* et *le prsident de la Commission Europenne* refrent-elles la mme entit?

Le point sur les tches

- **Rsolution de co-rfrence**:
  au sein d'un mme document, identifier que *Frdric Mitterrand*, *Mitterrand*, *FM* ont le mme rfrent (quel qu'il soit)
- **Clustering de mentions**:
  pour une collection de documents, identifier que *Frdric Mitterrand*, *Mitterrand*, *FM* ont le mme rfrent (avec ou sans rfrentiel)
- **Liaison d'entits**:
  tant donns des documents, identifier les mentions d'entits et lier chacune d'elles un rfrent d'une base de connaissances

Les bases de rfrents

- forte utilisation de Wikipdia (‘*wikification*’) et/ou DBpedia
- lorsque le rfrent d’une mention est absent de la base $\rightarrow$ NIL
  Non trivial: possibilits de mentions dont le rfrent est absent de la base, mais dont un homonyme y est prsent.
  $\rightarrow$ vers la population de bases de connaissances (cf. TAC-KBP)

Formalisation

Etant donns:

- l'ensemble des mentions dans des textes
- l'ensemble des entits rfrences dans une base

la liaison est une application depuis le domaine des mentions vers le domaine des rfrents.

- ni injective: plusieurs mentions peuvent tre associs un rfrent
- ni surjective: tous les rfrents n'ont pas tre lis

La liaison peut s'appuyer sur une *reconnaissance* des EN : restreindre les rfrents potentiels selon le type facilite les choses (e.g. *Washington* pour le gouvernement des tats-Unis).

Formalisation

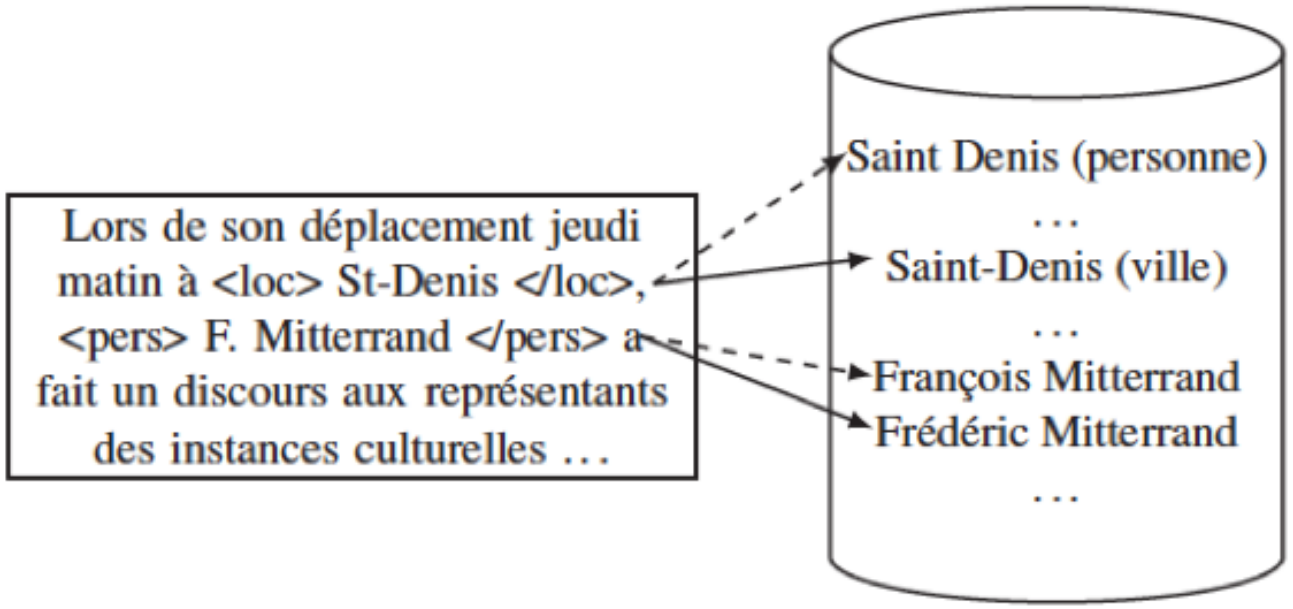


Figure 5.1. *Mentions du texte et références de la base de connaissances*

Etapes du processus de liason

1. Recherche des mentions d'EN dans les textes
2. Selection de candidats dans la base
3. Liaison

Recherche des mentions

1. utilisation de systmes de reconnaissances d'EN (cf. section prcdente) exposs
2. large collection de mentions par simples heuristiques ou par lookup des noms de la base

N.B.: un collection 'grossire' de mentions n'est pas prjudiciable: l'tape suivante fait un filtre.

Selection de candidats

Confronter chaque mention tous les rfrents est coteux et peu efficace.

$\rightarrow$ on slectionne, pour chaque mention, les rfrents *candidats* la liaison.

Approche:

- prise en compte des mots des mentions du ct du texte
  (ou les formes de surface)
- prise en compte des variantes (ou *dnominations, lexicalisations*)
  prsentes dans la base pour les rfrents.

Rcupration de dnominations dans la base

Pour Wikipedia:

Iment gnrique	**Donne Wikipdia**
Nom du rfrent	Titre de l'article
Autres appellations indiques	Texte en gras du premier paragraphe
Synonymes	Pages de redirection
Textes des pointeurs	Liens internes

Table 1: Extraction de variantes des rfrents dans une ressource

Selection de candidats

Donc: Pour une mention donne, rcupration des rfrents qui ont une variante correspondant aux mots de la mention.

Exemple: la mention *F. Mitterrand* aura comme candidats toutes les personnes dont le prnom commence par un *F* et dont le nom de famille est *Mitterrand*

Selection de candidats

Traitements de surface pour ne pas tre trop contraignant:

- insensibilit la casse (majuscules / minuscules) ;
- variantes lies aux jeux de caractres (diacritiques, liaisons, ponctuations, etc.) ;
- suppression des mots-outils ;
- suppression d'lments entre parenthses ;
- gnration automatique d'acronymes partir des formes de surface ;
- etc.

Enjeu: multiplier les variantes linguistiques pour chaque rfrent afin de ne pas manquer des mentions.

Remarque

Les 2 tapes de collection de mentions et de slection de candidats peuvent tre remplaces par une reconnaissance d'EN performante privilgiant le rappel sur la prcision.

Liaison

Etape la plus importante.

Objectif: associer chaque mention le candidat le plus vraisemblable (ou NIL)

Mthode: prises en compte d'indices du ct de la mention et du ct du rfrent, et calcul d'une distance.

Indices

- Cot mentions:
  - mots de la mention
  - contexte immdiat de la mention
  - texte du document
- Cot rfrents:
  - lments textuels de sa description (titre, synonymes, rsum, article)
  - autres proprits attribues au rfrent (infobox)
  - entits et concepts associes (internes et externes)

  Beaucoup d'informations disponibles, il faut choisir les plus pertinentes.

Calcul de similarit

- **Indices textuels**
  $\rightarrow$ possibilit de calculer une distance cosinus
  Dans quels cas cela marche-t-il bien?:
  - *Washington* LIEU vs. PERSONNE
  - *Karl Marx versus Thierry Marx*
  - *George H. W. Bush versus George W. Bush*
- **Indices structurels**:
  $\rightarrow$ slection du rfrent le plus populaire selon un critre.
  - pour tout rfrent, le nombre de liens pointant vers sa page
  - pour une ville, son nombre d'habitants

  Attention: non prise en compte des mentions: rsultats toujours identiques.

Performances

- **Premiers travaux exploratoires**
  briques lmentaires, dfinition des diffrentes sous-tches [**?**, **?**]
- **TAC 2009**
  - 82.2% d'accuracy (meilleur systme bas sur DBpedia)
  - dont 76.5% sur les entits lier et 86.4% pour les entits
- **TAC 2011**
  - 85% et 90% selon les donnes considres
- Performances des humains: autour de 90%

$\longrightarrow$ Bonnes performances sur des textes gnriques en anglais.

Rfrences:[**?**, **?**, **?**, **?**, **?**, **?**].

liaison:
à retenir

- tablissement de liens entre les textes et des bases de connaissances
- essor de la tches partir de 2007
- relatives bonnes performances pour l'anglais
- Wikipdia est la principale, voire unique, base de connaissances

Perspectives

Possibilit de s'appuyer sur les entits reconnues et lies pour mieux comprendre les textes
$\rightarrow$ bnficiaires: reconnaissance de la parole, la traduction automatique, le rsum automatique, recherche d'information

6. Evaluation

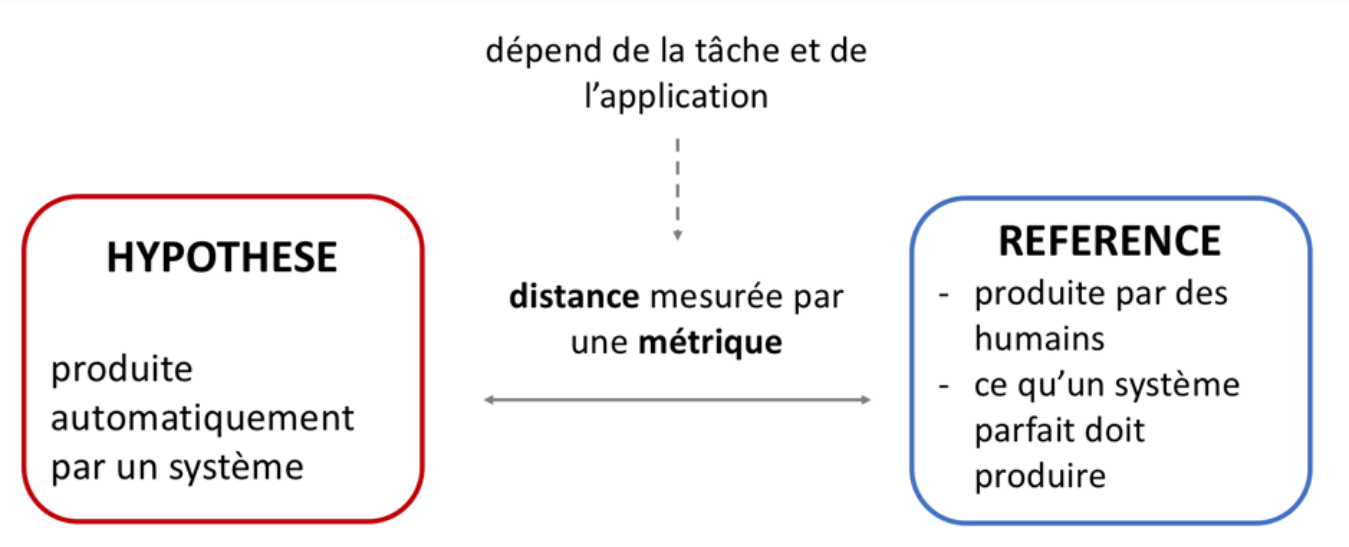
6. Evaluation

6.1 Introduction

Evaluer

- Premire formalisation de la procdure d'valuation: MUC3 [**?**]
- Motivation: avoir des des lments de comparaison stables et effectifs entre hypothses et rfrences

Protocole d'valuation



Objectif: mesurer quel point le systme trouve les "bonnes rponses"

Quelle "bonnes rponse"?

- traduction ou le rsum automatique : bonne rponses multiples
- EN: on peut supposer une seule et unique "bonne rponse"

Avantages

- **Transparence**: “rgles du jeu” connues par tous
- **Cot**: rduit par rapport une valuation manuelle pour chaque hypothse des systmes ;
- **Reproductibilit**: rutilisation au del des campagnes permettant une comparaison des rsultats dans la production scientifique

Ce qu'il faut pour valuer

1. Une mtrique mesurant la distance entre une rfrence et une hypothse ;
2. Un algorithme d'alignement de la rfrence et de l'hypothse.
3. Un algorithme de projection des entits annotes sur la transcription manuelle de rfrence vers la transcription automatique

6. Evaluation

6.2 Les mesures classiques

Les mesures classiques

Prcision

Ratio entre le nombre de rponses correctes et toutes les rponses donnes par un systme

$$P = \frac{C}{C + S + I} \tag{1}$$

- C : nombre d'objets **corrects** dans l'hypothse;
- I : nombre d'**insertions** par le systme ;
- S : nombre de **substitutions** par le systme (entits mal types).
- soit $C + S + I$: nombre total d'objets dans l'hypothse.

Rappel

Ratio entre le nombre de rponses correctes et le nombre des rponses attendues (i.e. prsentes dans la rfrence)

$$R = \frac{C}{C + S + D} \tag{2}$$

- D : nombre total d'**omission** (*deletions*) opres par le systme (entits non dtectes) ;
- $C + S + D$: nombre total d'objets dans la rfrence.

Exemple 1

REF: <pers> Bertrand Delanoë </pers> a été élu maire de <loc> Paris </loc>

HYP1 : <pers> Bertrand Delanoë </pers> a été élu <pers> maire </pers> de <loc> Paris </loc>

- Precision $= \frac{2}{3} = 0,67$
- Rappel $= \frac{2}{2} = 1$

$\rightarrow$ ici HYP1 produit du **bruit**

Exemple 2

REF: <pers> Bertrand Delanoë </pers> a été élu maire de <loc> Paris </loc>

HYP2: <pers> Bertrand Delanoë </pers> a été élu maire de Paris

- Precision $= 1$
- Rappel $= \frac{1}{2} = 0.5$

$\rightarrow$ HYP2 produit du **silence**

- La prcision tient compte des insertions et substitutions
- Le rappel tient compte des omissions

Comment combiner les 2 en une seule mesure?

F-mesure

F-mesure, dfinie comme la **moyenne harmonique entre Prcision et Rappel**:

$$F = (1 + \beta^2) \times \frac{P \times R}{\beta^2 P + R} \tag{3}$$

O β est un **poids** permettant d'ajuster l'importance de P ou R (si 1, gale importance).

Exemples

REF: <pers> Bertrand Delanoë </pers> a été élu maire de <loc> Paris </loc>

HYP1 : <pers> Bertrand Delanoë </pers> a été élu <pers> maire </pers> de <loc> Paris </loc>

HYP2: <pers> Bertrand Delanoë </pers> a été élu maire de Paris

$$F(HYP1) = (1 + 1^2) \times \frac{0,67 \times 1}{1^2 \times 0,67 + 1} = 0,80 \quad (4)$$

$$F(HYP2) = (1 + 1^2) \times \frac{1 \times 0,5}{1^2 \times 1 + 0,5} = 0,67 \quad (5)$$

Inconvnients des mesures classiques

- Fusionner P et R minimise le poids des erreurs d'insertion et d'omission par rapport aux erreurs de substitution, quel que soit β [**?**]
- Avec les typologies fines et complexes, besoin d'une mtrique diffrenciant les erreurs.

Diffrents types d'erreur

REF: the <pers.ind> president of Ford </pers.ind>

HYP1 : the <pers.ind> president </pers.ind> of Ford
→ erreur de frontire

HYP2 : the <pers.coll> president of Ford </pers.coll>
→ erreur de sous-type

HYP3 : the <pers.coll> president </pers.coll> of Ford
→ erreur de sous-type et de frontire.

6. Evaluation

6.3 Les mesures bases sur le dcompte d'erreurs

Mesures bases sur le dcompte d'erreurs: ERR

ERR, Error Per Response

- dfinie lors de MUC5 [**?**]
- inspire du taux d'erreurs mots (WER pour *Word Error Rate*) en RAP [**?**]
- mesure des erreurs: plus le taux est bas, mieux c'est.

$$ERR = \frac{S + D + I}{C + S + D + I} \tag{6}$$

ERR: exemples

REF: <pers> Bertrand Delanoë </pers> a été élu maire de <loc> Paris </loc>

HYP1: <pers> Bertrand Delanoë </pers> a été élu <pers> maire </pers> de <loc> Paris </loc>

HYP2: <pers> Bertrand Delanoë </pers> a été élu maire de Paris

$$ERR(HYP1) = \frac{0+0+1}{2+0+0+1} = \frac{1}{3}$$

$$ERR(HYP2) = \frac{0+1+0}{1+0+1+0} = \frac{1}{3}$$

ERR: problme

Le poids des insertions est moins important que celui des substitutions et des omissions[**?**].

Une augmentation de I provoque une augmentation de ERR moins importante qu'une augmentation de $S + D$.

$$ERR = \frac{S + D + I}{N + I} \tag{7}$$

Avec $N =$ nombre d'entits dans la rfrence.

ERR: problme

Pour $N = 100$, $S + D = 10$, $I = 10$, on a:

$$ERR = \frac{10 + 10}{100 + 10} = \frac{20}{110}$$

Si on augmente $S + D$ de 10:

$$ERR = \frac{20 + 10}{100 + 10} = \frac{30}{110} = \mathbf{0{,}27}$$

Si on augmente I de 10:

$$ERR = \frac{10 + 20}{100 + 20} = \frac{30}{120} = \mathbf{0{,}25}$$

De plus, avoir I dans le dnominateur rend les rsultats non comparables.

Mesures bases sur le dcompte d'erreurs: SER

SER: Slor Error Rate

- propose par [**?**]
- identique au WER utilis en RAP
- utilise lors de ACE, ESTER-2, QUAERO et ETAPE
- suppression du nombre d'insertion (I) du dnominateur:

$$SER = \frac{S + D + I}{C + D + S} = \frac{S + D + I}{R} \tag{8}$$

o R = nombre total d'entits de la rfrence.

SER

Possibilit d'affiner l'importance relative des erreurs:

$$SER = \frac{\alpha_1 S_t + \alpha_2 S_f + \beta D + \gamma I}{R} \tag{9}$$

- S_t et S_f : nombre total de substitution de type et de frontires ;
- D et I: nombre total d'omissions et insertions ;
- $\alpha_1\, \alpha_2\, \beta$ *et* γ : poids affectes chaque catgorie d'erreur.

SER: inadapt aux imbrications

- reprsentation en “slot” des hypothses et de la rfrence
- slot= un segment de texte avec des frontires (dbut/fin) et un type

$\longrightarrow$ structure plate qui ne peut pas traiter les entits imbriques

Mesures bases sur le dcompte d'erreurs: ETER

ETER: Entity Tree Error Rate

Base sur une comparaison des arbres d'entits [?]

$$ETER = \frac{I + D + \sum_{(e_r, e_h)} E(e_r, e_h)}{N_E} \tag{10}$$

- I : nombre total d'insertions d'arbre-entit ;
- D : nombre total d'omission d'arbre-entit ;
- (e_r, e_h) : paires d'entits-arbres rfrence/hypothse associes l'issue de l'alignement ;
- $E(r, h)$: erreur calcule pour chaque paire d'entit-arbre (e_r, e_h) (peut tre zro) ;
- N_E: nombre d'entit-arbre dans la rfrence.

ETER

En haut, un alignement bas sur les slots, en bas le mme bas sur les arbres d'entits [**?**]

ETER

$$ETER = \frac{I + D + \sum_{(e_r, e_h)} E(e_r, e_h)}{N_E} \tag{11}$$

Le calcul d'erreur pour les paires d'entit-arbre $E(r, h)$ a 2 parties

- erreur de dtection et de classification de l'entit
- erreur de dcomposition E_c

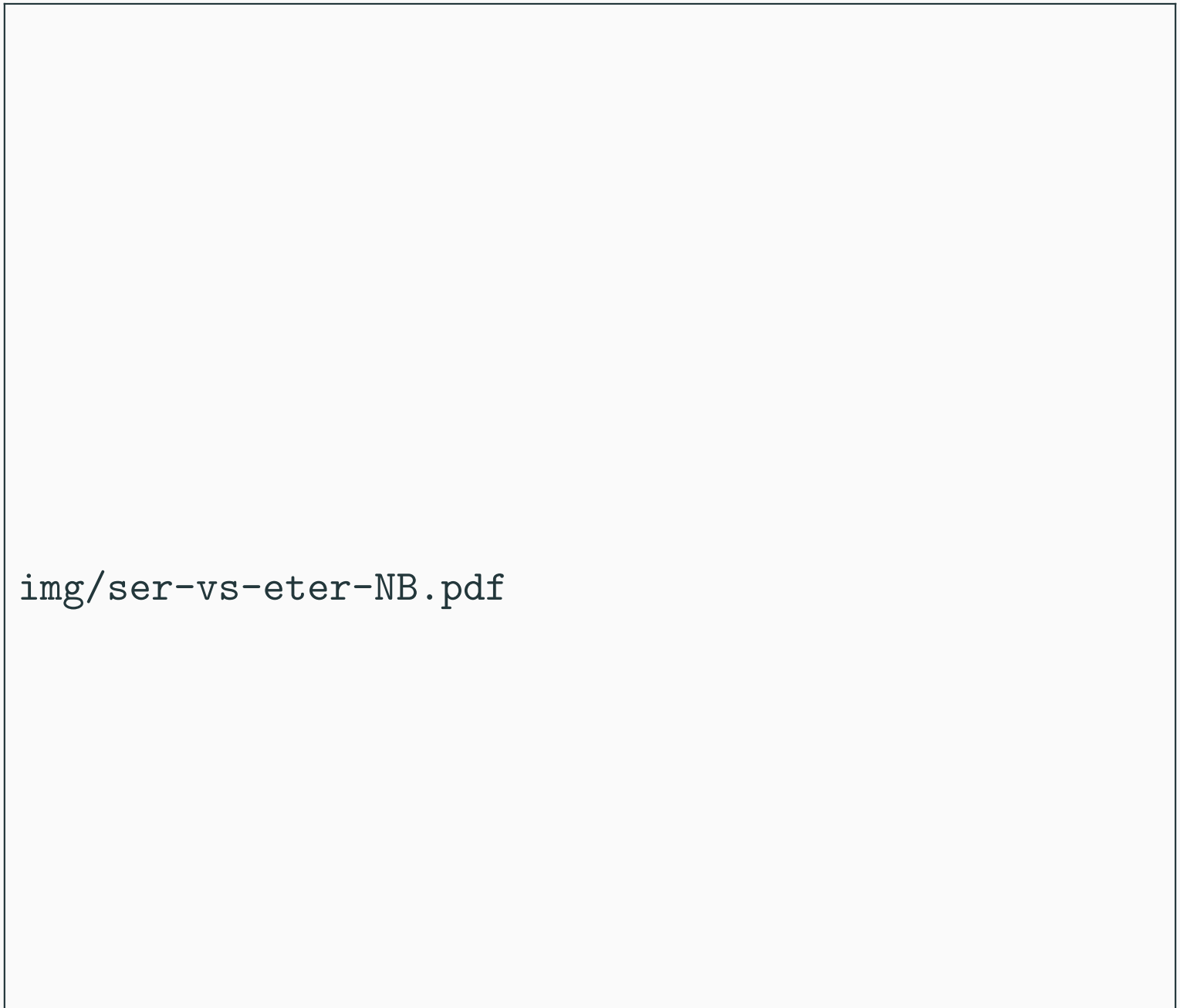
ETER: erreur de dcomposition

$$E(r,h) = (1-\alpha)E_T(e_r,e_h) + \alpha E_c(e_r,e_h), \alpha \in [0..1] \qquad (12)$$

- $E_T(e_r,e_h)$: erreur de classification, dpend de la distance entre (e_r,e_h) ;
- $E_c(e_r,e_h)$: erreur de dcomposition, dpend de la distance entre les constituants des entits-arbres (e_r,e_h) ;
- α paramtre fixant le poids relatif de la dcomposition par rapport la classification.

$\rightarrow$ $E_c(e_r,e_h)$ se rapproche d'un *SER* local

ETER: exemple



6. Evaluation

6.4 Zoom sur donnes de parole

Corpus et campagnes d'valuation

- Assez peu de corpus et de campagnes d'valuation
  - en France : ESTER 1 et 2, ETAPE (+ QUAERO), REPERE (pour les personnes, multimodal)
  - l'international : campagne ACE (2000-2008)
- Difficile de comparer REN sur textes et REN sur parole car on ne dispose pas de corpus et campagnes comparables (types de donnes + typologies)
- Rsultats nettement diffrents entre transcriptions manuelles et transcriptions automatiques

REN sur transcriptions automatiques et manuelles

Pour comparer simplement, utilise par [**?**] :

$$PAE(e) = 100 * \frac{NB_A(e) - NB_M(e)}{NB_M(e)} \tag{13}$$

Avec :

- e une erreur de REN de type omission, insertion ou substitutions;
- NB_A le nombre des erreurs de type e sur les transcriptions automatiques;
- NB_M le nombre des erreurs de type e sur les transcriptions manuelles.

REN sur transcriptions automatiques et manuelles

PAE ETAPE-1

REN sur transcriptions automatiques et manuelles

PAE ETAPE-2

REN sur transcriptions automatiques et manuelles

Quelques constats

1. sur ETAPE plus un systme ASR insre de mots, plus un systme REN insre d'entits (pas observes sur les donnes QUAERO)
2. impact fort des erreurs (notamment omissions et insertions) sur la mention de l'entit
3. impact non nul des erreurs des mots qui introduisent une EN

Point 2 et 3 en lien direct avec la faon dont les systmes REN sont dvelopps.

Contacts

Maud Ehrmann
EPFL-DHLAB
maud.ehrmann@epfl.ch

Matteo Romanello
LIMSI
matteo.romanello@epfl.ch

Simon Clematide
ICL, Zurich
simon.clematide@uzh.ch

References i