Named Entity Processing for Historical Texts

DEFINITION, RESOURCES, APPROACHES, APPLICATIONS

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

9 July 2019

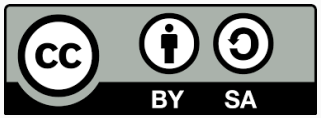
[1] EPFL-DHLAB, Lausanne, Switzerland

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

Slides

- **Available at**
  https://github.com/impresso/named-entity-tutorial-dh2019

- **License**: CC-BY-SA 4.0



Objectives

- Getting to know about:
  - the origins of named entity processing
  - the resources needed for their processing
  - the evaluation protocols
  - the tools and algorithms used for their recognition, classification and disambiguation.
- Getting your hands on:
  - rule-based NERC system
  - two deep learning-based NERC systems (spacy, flair)
  - entity linking with Babelfy and Aida

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 We will use mentimeters to do interactive Q&A sessions.

Introduction to Named Entity Processing: Outline

1. Context and Applications
2. 'Definition'
3. Resources
4. Recognition and classification
5. Linking
6. Evaluation
7. Zoom on DH

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

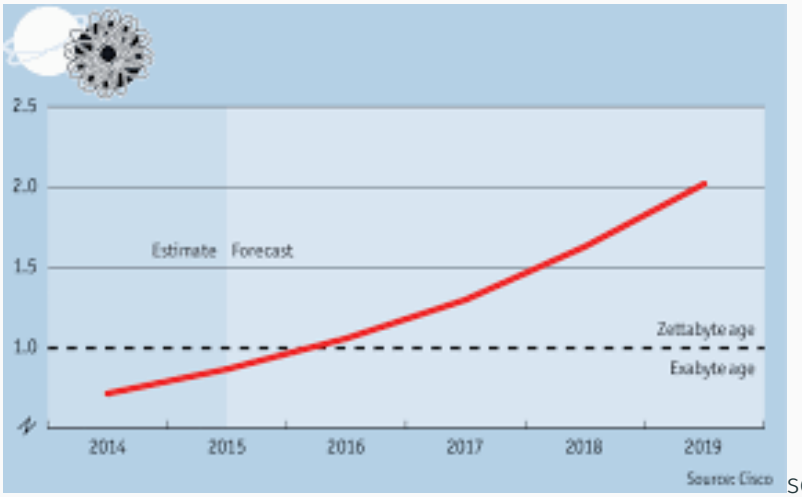
What: EVERYTHING

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:

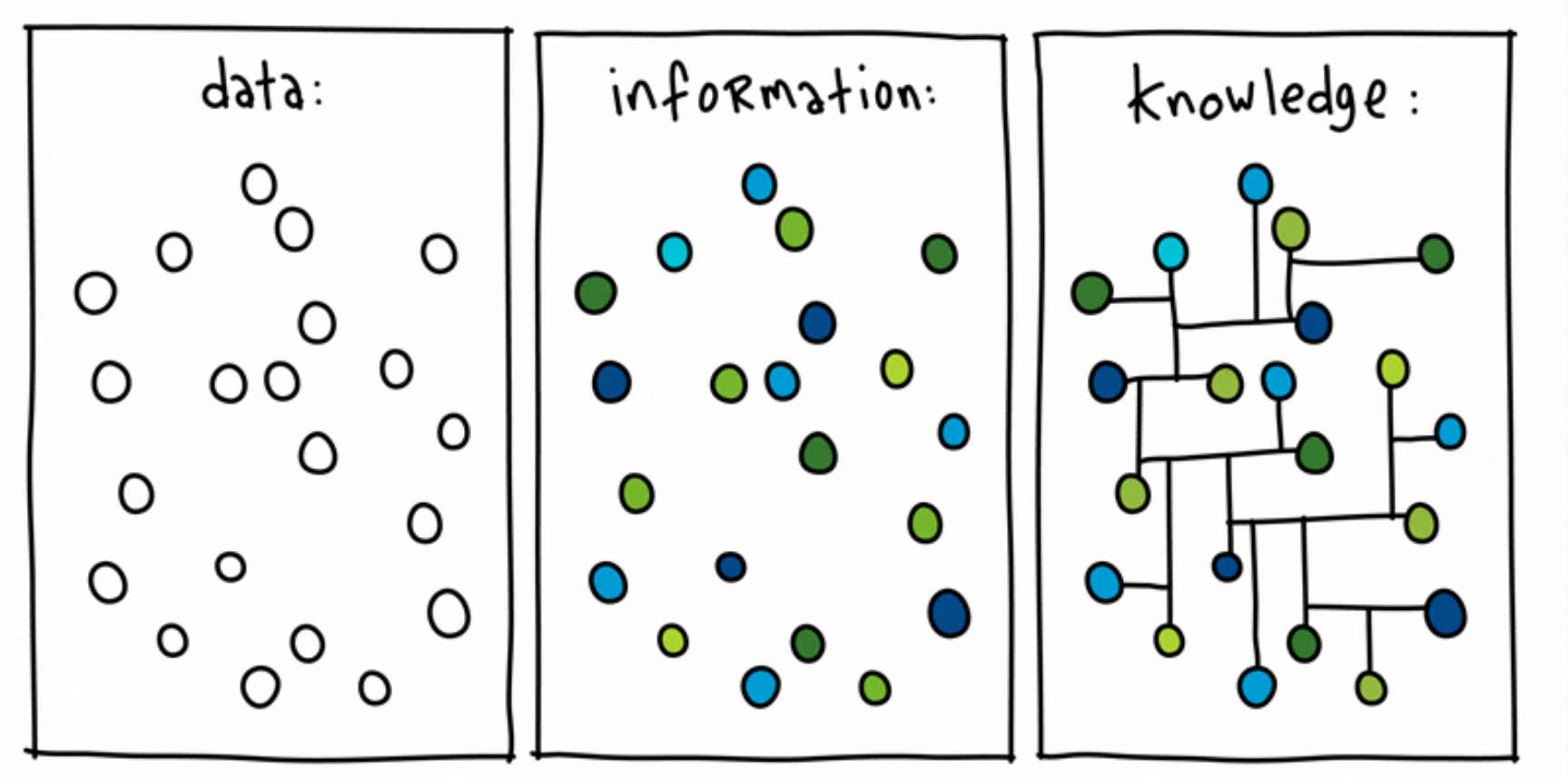
- storage more and more expensive
- above all: data exploitation, extract useful information

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-à-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

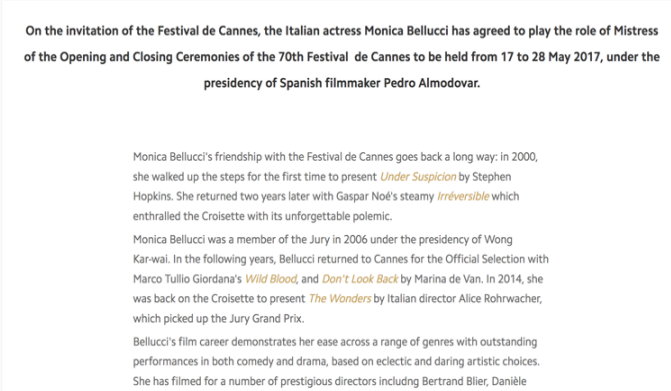
Follow

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 Information Extraction

- **Named entity processing**:
  recognition, categorization and disambiguation
  - *Monica Belluci* and *Pedro Almodovar* are PERSON.
  - *Monica Belluci* $\xrightarrow{\textit{reference}}$ `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* $\xrightarrow{\textit{instance\_of}}$
    `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 (1/2)

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

→ Today, generic approach to text comprehension is still an utopia

From text understanding to Information Extraction (2/2)

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)

Example of 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

Given a document, participants were asked to:

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

Evolution of MUC conferences

- **1987 - MUC-1:** no specific task, military reports;
- **1989 - MUC-2:** templates with 10 slots; definition of evaluation measures (precision and recall).
- **1991 - MUC-3:** news reports on terrorist events; 18 slots.
- **1992 - MUC-4** idem, 24 slots;
- **1993 - MUC-5** more complex tasks, out-of-domain material, 2 languages, 11 templates with 48 hierarchical slots
- **1995 - MUC-6:** reformulation of objectives and definition of subtasks;
- **1997 - MUC-7:** continuation

Evolution of MUC conferences

MUC-6: reformulation of objectives and definition of sub-tasks

- **Independent technologies and components**
  → Definition of the notion of ”named entity”
  → Definition of the task of recognition and classification of NE

- **Portable systems**
  → definition of generic templates

- **Consideration of ”basics building blocks of understanding”**
  → co-reference, word sense disambiguation,
  predicate-argument structure, etc.

Subsequent evaluation campaigns

- MET, IREX, CONLL, ACE, ESTER, ETAPE, HAREM, EVALITA, GERMEVAL, TREC, TAC, etc.
- All using predefined templates

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

Objective: discover information about NEs as found in a large corpus and incorporate this information into a knowledge base.

Given an entity, one should find its attributes; e.g. for PERS:

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

$\longrightarrow$ Back to text understanding !

TAC - KBP

The task remains **very complex** despite significant progress:

- **2010**: the best system did not exceed 0.30 F-measure
- **2014**: highest score was 0.36

[SJ14]
`https://tac.nist.gov`

Back to text understanding

Team	NER			NERC			NERLC			KBIDs			CEAFmC+		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
Tri-lingual															
5	**83.2**	**67.3**	**74.4**	**76.8**	**62.2**	**68.8**	**62.6**	**50.7**	**56.0**	**73.1**	**64.9**	**68.8**	**60.7**	**49.1**	**54.3**
18	52.8	54.8	53.8	29.8	30.9	30.3	22.6	23.4	23.0	64.1	46.9	54.2	19.7	20.5	20.1
16	81.7	53.0	64.3	71.7	46.5	56.4	5.5	3.5	4.3	0.0	0.0	0.0	4.8	3.1	3.7
Chinese															
5	84.8	62.9	**72.2**	79.6	59.1	**67.8**	**65.1**	**48.3**	**55.4**	79.9	**64.9**	**71.7**	**64.0**	**47.5**	**54.5**
14	75.0	60.5	67.0	70.0	56.5	62.6	47.8	38.5	42.7	**84.4**	38.7	53.1	46.3	37.4	41.4
18	68.2	47.4	55.9	38.8	26.9	31.8	31.5	21.9	25.8	62.3	44.4	51.8	30.6	21.3	25.1
15	79.8	56.2	66.0	73.9	52.0	61.1	14.7	10.3	12.1	0.0	0.0	0.0	13.9	9.8	11.5
20	56.2	**71.5**	63.0	51.7	**65.9**	57.9	9.9	12.7	11.1	0.0	0.0	0.0	8.9	11.4	10.0
16	**85.4**	50.8	63.7	**81.1**	48.3	60.5	5.0	3.0	3.7	0.0	0.0	0.0	4.6	2.8	3.5
English															
5	77.5	66.7	71.7	71.5	61.5	66.1	**57.9**	49.8	**53.5**	63.6	**68.2**	**65.8**	**54.1**	46.5	**50.1**
14	78.6	79.1	**78.8**	72.6	**73.0**	**72.8**	52.9	**53.2**	53.0	**70.4**	49.8	58.4	48.8	**49.1**	49.0
15	73.0	**79.5**	76.1	66.1	71.9	68.9	23.2	25.3	24.2	0.0	0.0	0.0	21.1	22.9	22.0
21	**90.8**	62.5	74.1	**83.3**	57.3	67.9	26.9	18.5	21.9	0.0	0.0	0.0	23.5	16.2	19.2
18	55.9	70.5	62.4	31.7	39.9	35.3	19.5	24.6	21.8	66.9	50.5	57.6	16.0	20.2	17.9
16	78.5	48.9	60.3	71.3	44.5	54.8	7.8	4.9	6.0	0.0	0.0	0.0	7.0	4.4	5.4
24	51.5	32.9	40.1	29.7	19.0	23.2	5.2	3.3	4.0	0.0	0.0	0.0	4.9	3.1	3.8
Spanish															
5	**86.6**	**74.3**	**80.0**	**78.5**	**67.4**	**72.5**	**64.1**	**55.0**	**59.2**	**76.4**	**62.1**	**68.5**	**62.8**	**53.9**	**58.0**
18	40.9	50.4	45.1	22.7	28.0	25.1	19.9	24.6	22.0	64.0	46.6	53.9	16.2	20.0	17.9
16	84.9	58.7	69.4	63.5	43.9	51.9	5.2	3.6	4.2	0.0	0.0	0.0	4.5	3.1	3.7

Table 5: Overall Tri-lingual Entity Discovery and Linking Performance (%) during the First Evaluation Window.

[JPZ+17]

Open information extraction

- **New extraction paradigm** defined in 2007 by [BCS+07]
- Extract **all information and relations** found in texts in the form of tuples

Example:
If he wins five key states, Republican candidate Mitt Romney will be elected President in 2008
$\rightarrow$ (Republican candidate Mitt Romney; will be elected President in; 2008)

And in Digital Humanities?

- Same explosion of information due to massive digitization
- Information also hidden in (historical) 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

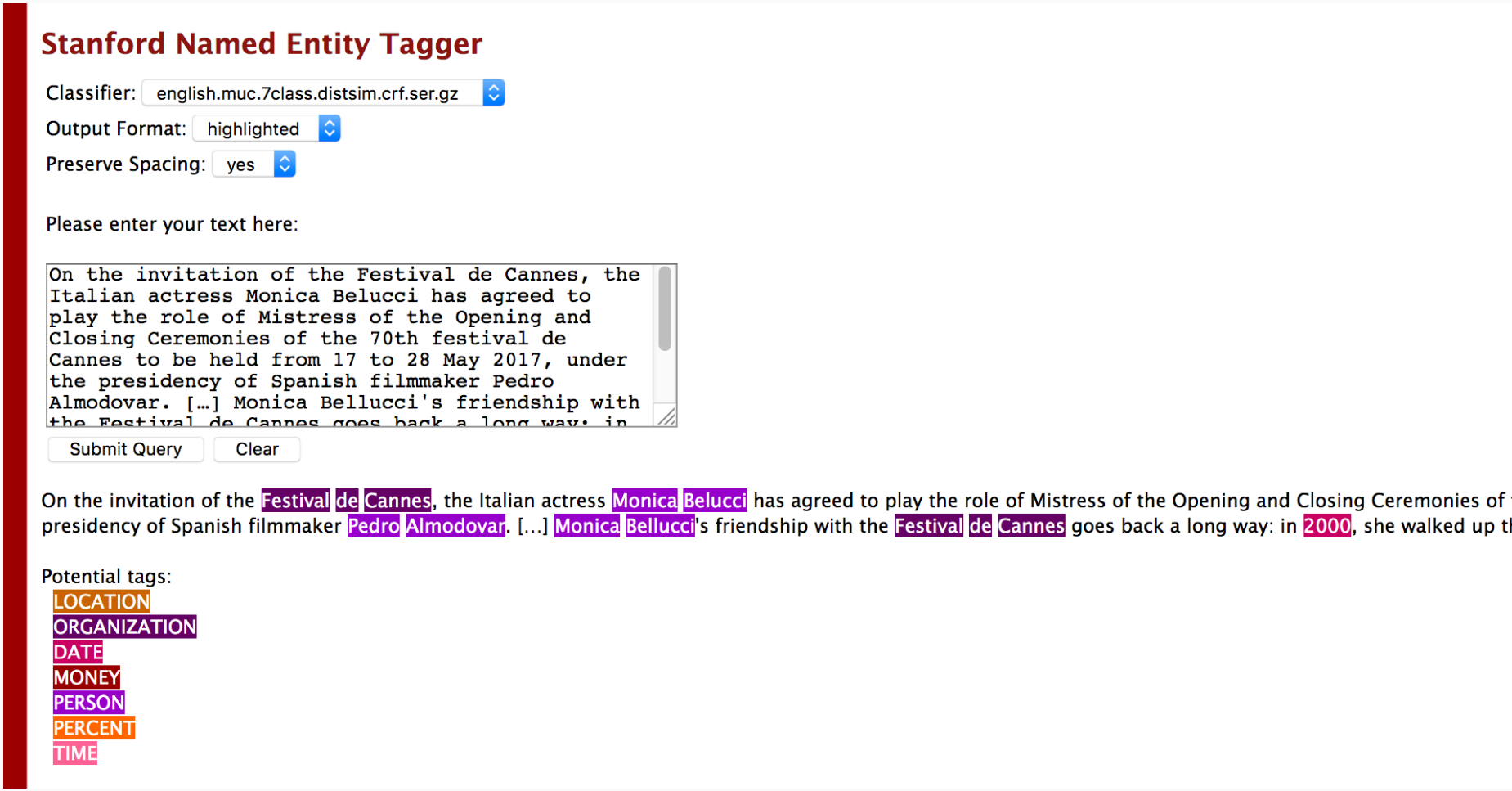
On the invitation of the Festival de Cannes, the Italian actress Monica Belucci has agreed to play the role of Mistress of the Opening and Closing Ceremonies of the 70th festival de Cannes to be held from 17 to 28 May 2017, under the presidency of Spanish filmmaker Pedro Almodovar. [...] Monica Bellucci's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

Stanford NER



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.

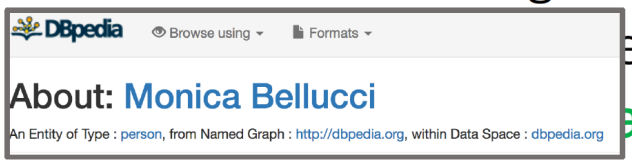
PERSON, ORGANIZATION, LOCATION, DATE

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.*

Examples from [BH04]

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. $\rightarrow$ leave = ”go away from a place” (#1 WordNet)

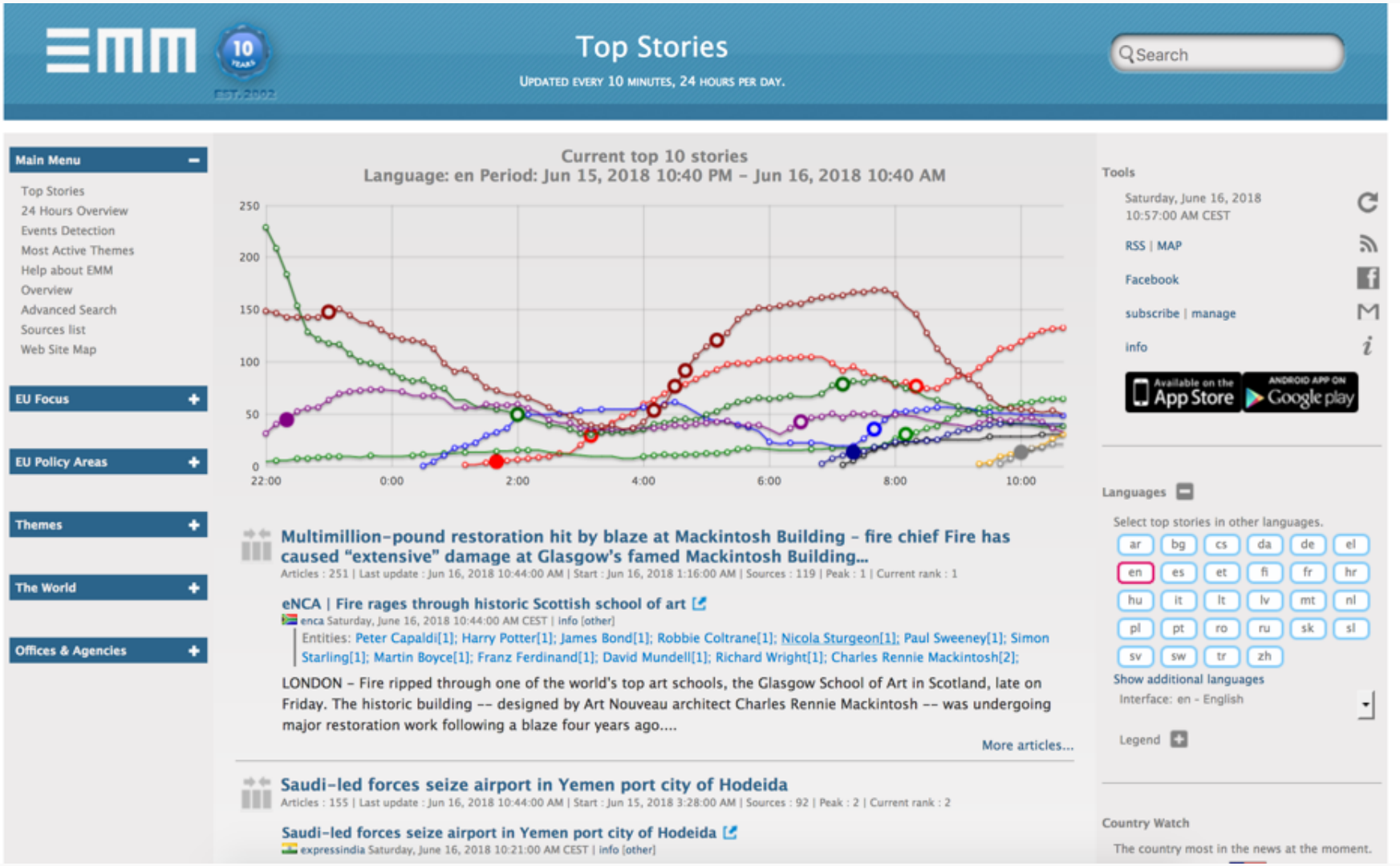
  Some wonder if they will leave the Socialist party. $\rightarrow$ leave = ”remove oneself from an association with or participation in” (#8 WordNet)

‘Real world’ applications

Information extraction and media monitoring

- Knowledge base population
- Alerts on certain topics or entities

European Media Monitor



http://emm.newsbrief.eu

Europena Media Monitor

‘Real world’ applications

- **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
- **Anonymization**
- **Various text analytics**

What about you?

Interactive session Go to mentimeters.

Context: take-home message

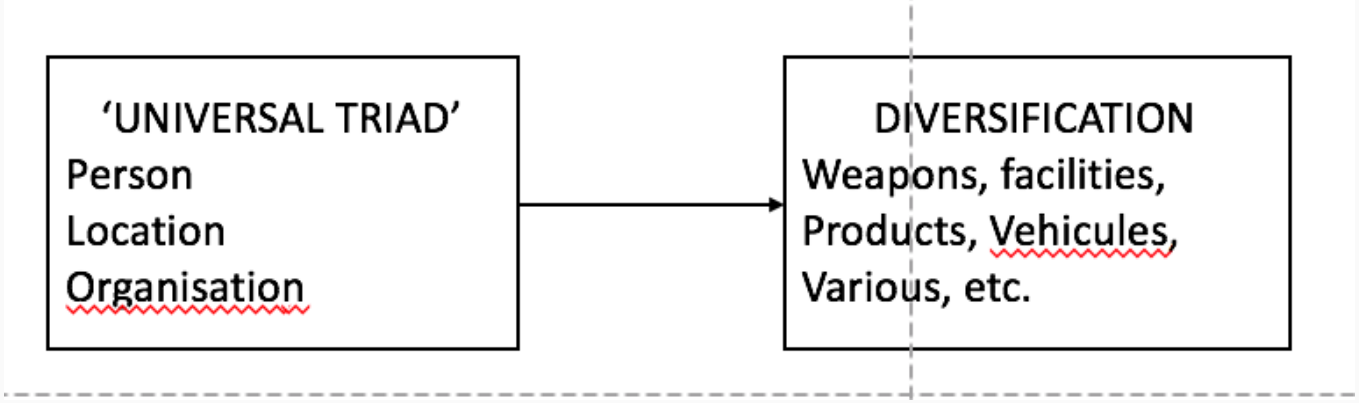
- The concept of NE appeared in the 90s during evaluation campaigns on text understanding.
- NEs quickly gained prominence and became one of the central hubs in automated text analysis systems.

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, Abbé 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 went bankrupt.

- Metonymy

  *The **BMW** slowed down. **France** wants to keep the head of IMF. The journalist is reporting from the **UN**.*

- “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

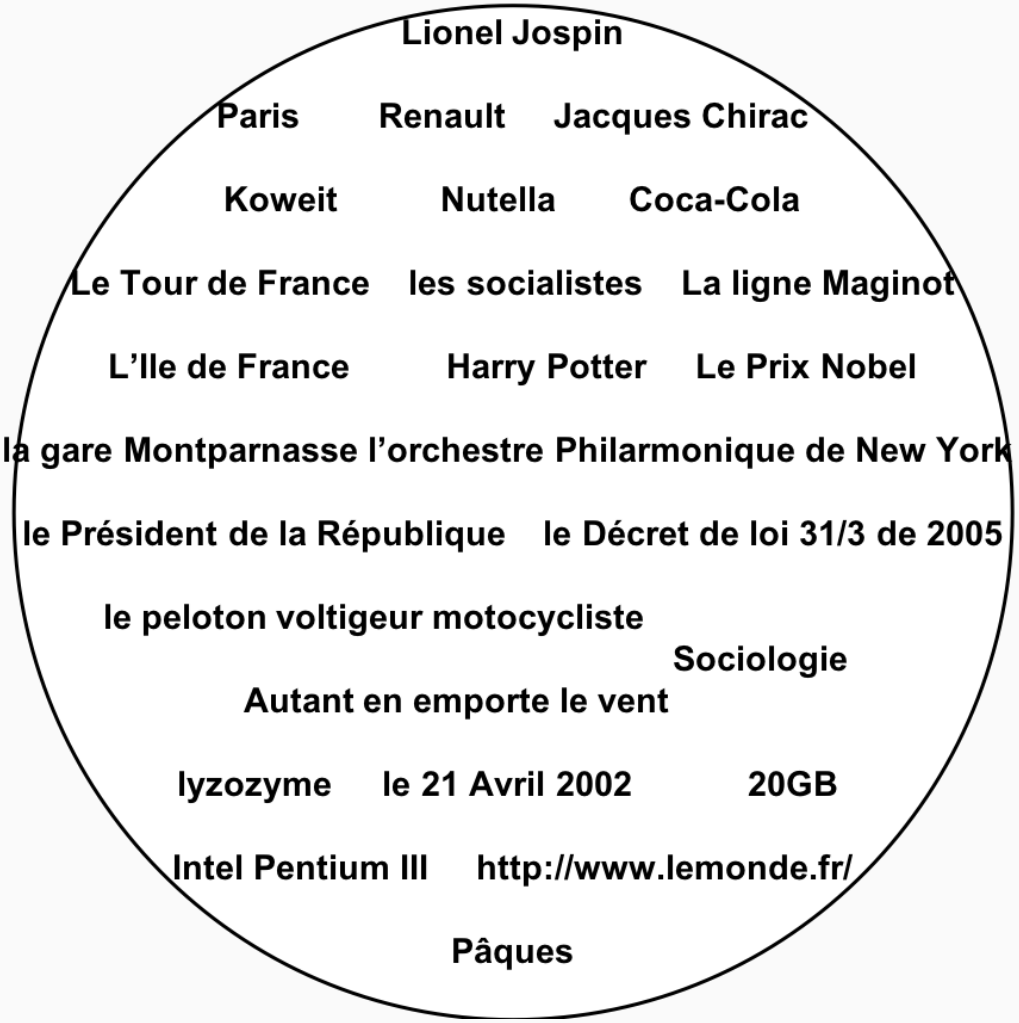
Heterogeneity of achievements

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

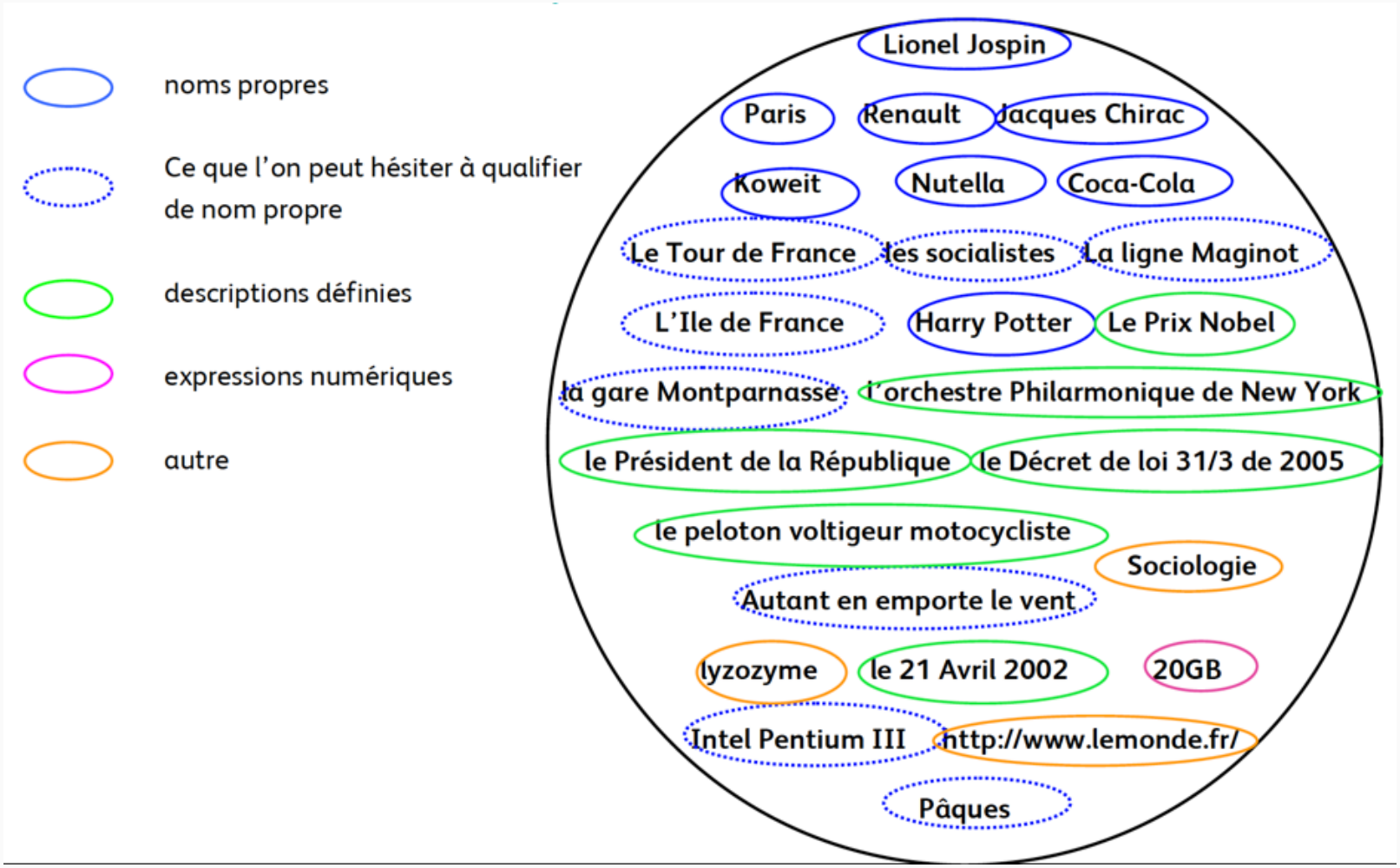
$\longrightarrow$ Question : What are named entities ?

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. [Ehr08]

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
  - **identity**: an individual considered as a recognizable whole at all time.
- Definite descriptions
  - **presupposition** of existence and uniqueness

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

→ Proper names and definite descriptions refer to unique entities.

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
  the president, the President of the French Republic in 2003

Linguistic characterisation

- **NEs do not correspond to one linguistic category**
  'More than proper names, but less than definite descriptions'.
- **We can only specify of a 'behavior'**
  - reference to a unique entity
  - referential autonomy

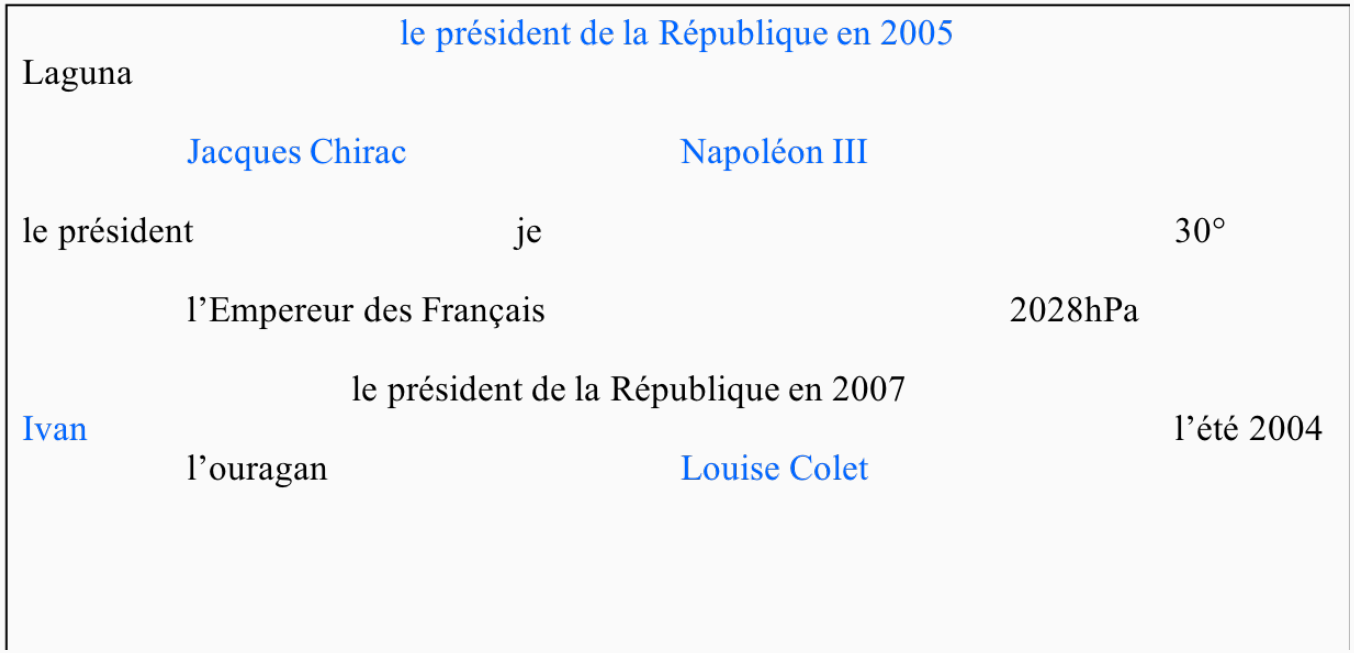
Jacques Chirac, the President of the Republic, the blue suit of the president
$\longrightarrow$ the linguistic perspective is not sufficient

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

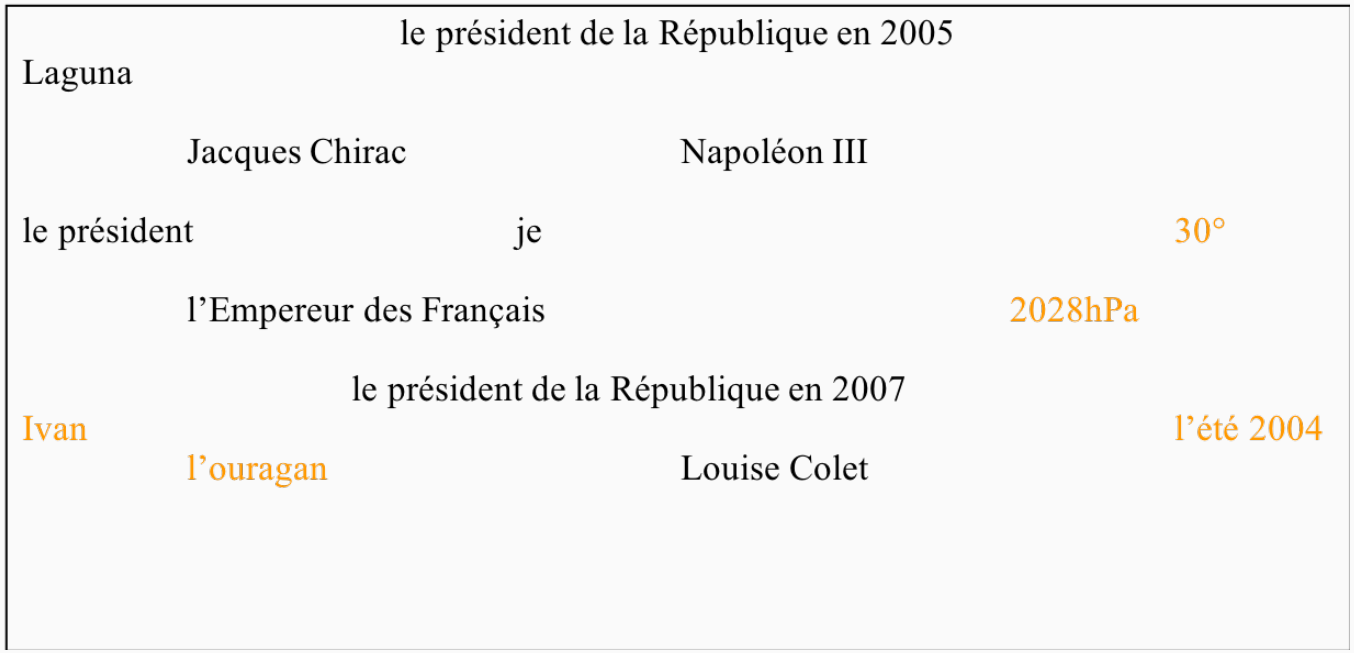
Given an application model and a corpus, we call named entity any linguistic expression that refers to a unique entity of the model autonomously in the corpus.



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

Illustration

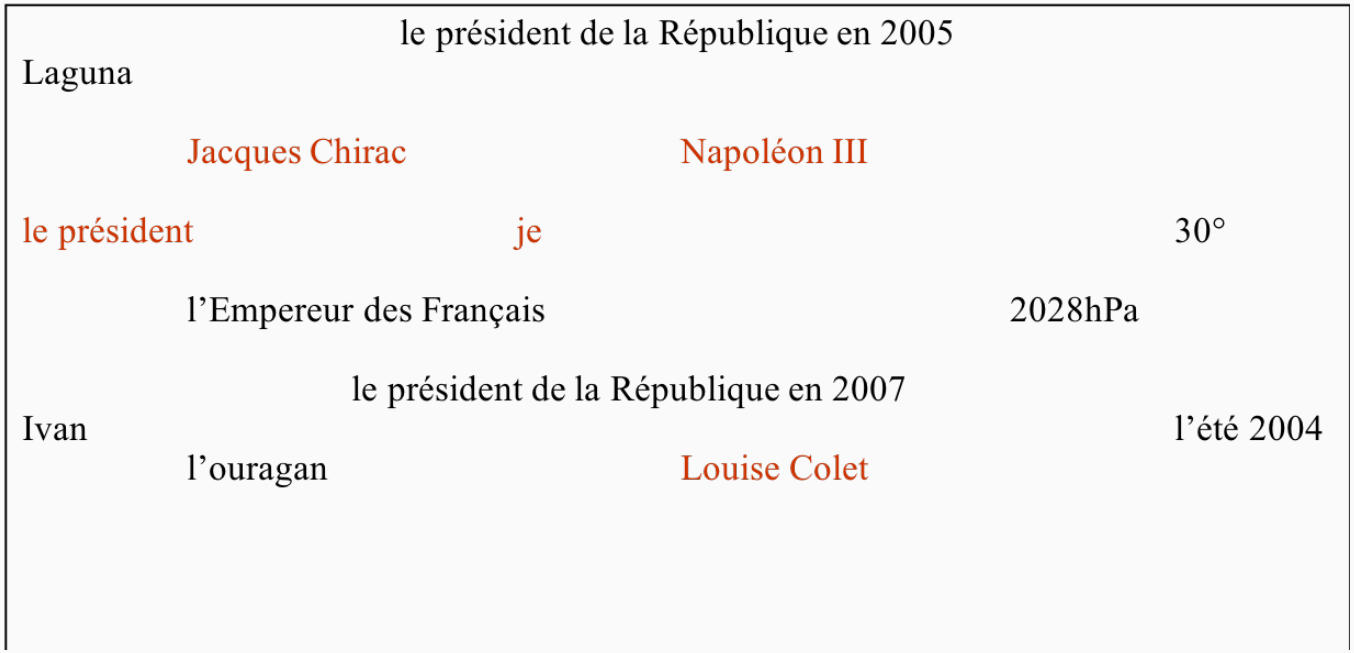
Given an application model and a corpus, we call named entity any linguistic expression that refers to a unique entity of the model autonomously in the corpus.



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’

No clear-cut definition, but a theoretical framework:

- 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 an application model.

Named entities: an NLP ‘creation’

Consequences

- do expect heterogeneity and variability of the set ’named entities’
- do not expect regularities in lexical units forming NEs
- from a methodological viewpoint: imperative need to explain the model to support decision criteria to annotate

Which entities would you annotate?

When the German chancellor was asked this week why she would not railroad Italy and the so-called Visegrád group of countries – Poland, the Czech Republic, Slovakia, Hungary – into accepting the former Dutch foreign minister Frans Timmermans, a critic of populist governments, as European commission president, Angela Merkel's answer was telling.

"The Brexit is looming on the horizon," Merkel said in reference to the need to avoid tensions when appointing the next head of the commission. "Other important issues are on the table. I think we need to treat each other with care."

`https://www.theguardian.com/world/2019/jul/04/brexit-just-one-eu-headache-angela-merkel-avoids-rocking-boat`

NE definition: take-home message

- There exists a general agreement on the 'core' of the concept of NE, but periphery and details are much more vague.
- They do not correspond to a linguistic category, but their common point is a referential behavior.
- Do not be afraid of variability and hesitations.

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 (Person?) vspace0.1cm
  - along with a definition of their scope (is Asterix a Person?)

- Typologies differ according to domains and applications
  - different categories
  - different structures: flat or hierarchical
  - different category definitions

How to define a typology?

Approaches:

- top-down: categories derived from the application
- bottom-up: categories derived from the corpus
- 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 [SSN02] detailed its methodology to define a typology (200 categories!).

MUC typologies

- **Proper names** (ENAMEX): Location, Person, Organisation,
- **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)

→ **Flat, simple typology**

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 .

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**, **Object**, **Group**, **Other**
- ESTER-2: even more sub-types (e.g. pers.hum, pers.anim, loc.geo, loc.admin, etc) and consideration of nested entities

Refs: [TKSDM03, SSCV06, **?**]

Fine-grained characterisation

In 2009 the French **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

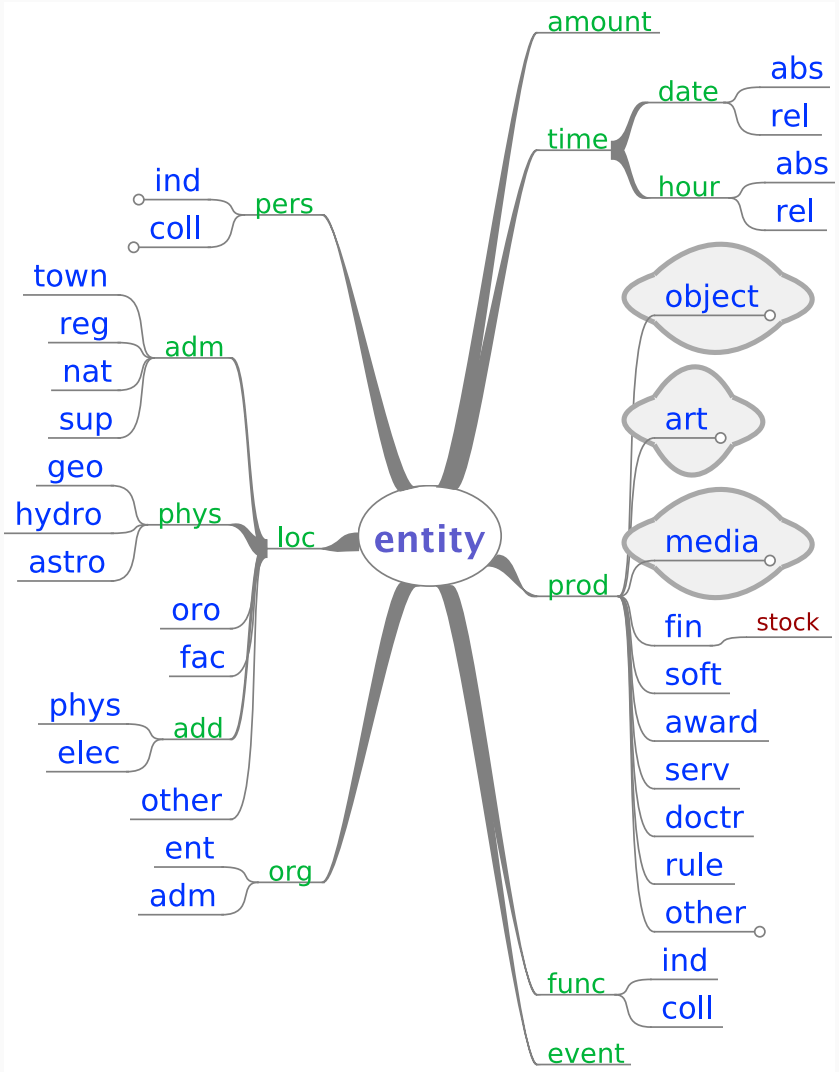
Ref: [GRZ+11, **?**]

Quaero typology

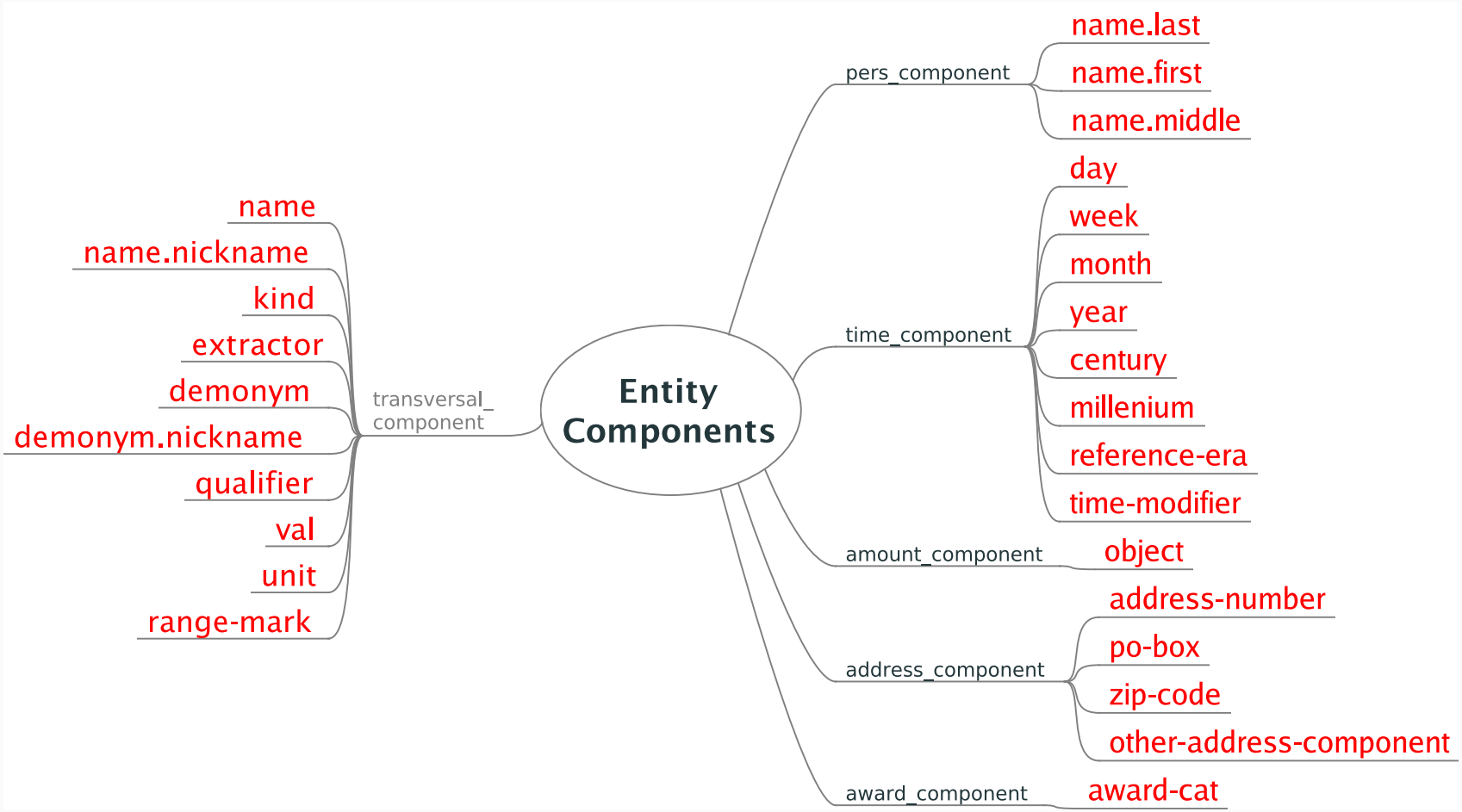
8 main categories

- **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;

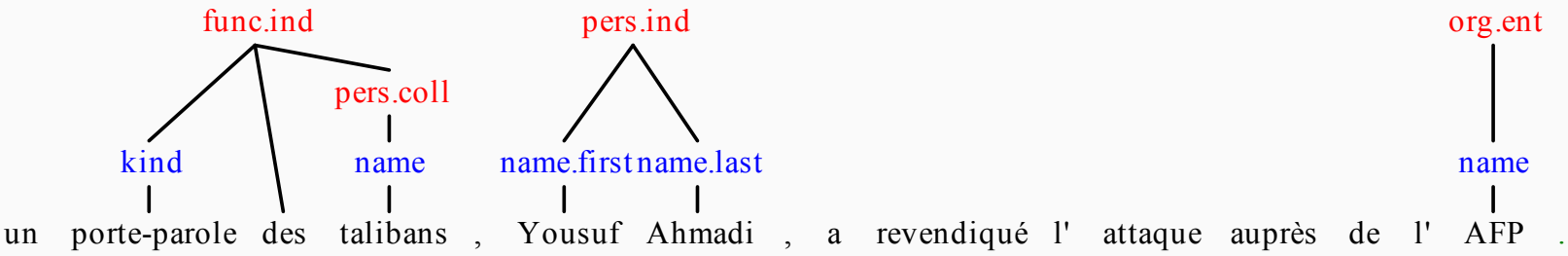
Quaero typology: sub-types



Quaero typology: components



Quaero typology: components



Comparison of typologies

MUC d'après le **Bureau du recensement des** LOC*[Etats-Unis]* , les revenus des ménages ont reculé pour la quatrième année consécutive en DATE*[2011]* .

ACE d'après le **ORG*[Bureau du recensement des Etats-Unis]*** , les revenus des ménages ont reculé pour la quatrième année consécutive en DATE*[2011]* .

ESTER d'après **le ORG*[Bureau du recensement des LOC[Etats-Unis]]*** , les revenus des ménages ont reculé pour la quatrième année consécutive en DATE*[2011]* .

QUA d'après **le ORG** ***[name [Bureau du recensement] des LOC [name[Etats-Unis]]]*** , les revenus des ménages ont reculé pour la quatrième année consécutive 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?

Typologies: 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.

Warning: time and resource consuming!

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

Inventory of existing NE annotated corpora

- `http://damien.nouvels.net/resourcesen/`
- Publication: Maud Ehrmann, Damien Nouvel, Sophie Rosset (2016). Named Entity Resources - Overview and Outlook, LREC.

Overview of existing (non DH) corpora

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

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

Corpora: 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

Gazetteers and Knowledge bases

Significant evolution of these resources since the 90s:

- **complexification**: basic 'gazetteers' $\longrightarrow$ bases encoding of more and more information.
- **less central**: essential for the recognition and classification of NEs up to now, moving to the background with deep learning.

Gazetteers (or 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 gazetteers

- 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 (‘prolexèmes’)
  - + their types
  - + their relations (e.g. synonymy, meronymy)
  - v2.2: French (100k units), English, Polish
- Developed by University François Rabelais of Tours, France

PROLEX

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

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

Conclusion on gazetteers

- Not all are published
- Initially: describe possible linguistic realizations of entities
- Evolution: enrichment according to 3 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'

Gazetteers 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 and classification

Quiz

What are the four main tasks in NE processing?

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 on *new texts*.

Requirements:

- quality: do not make too many mistakes (false positives)
- exhaustivity: do not miss too many entities (false negatives)
- robustness: do not fail in unseen, non-canonical, or noisy cases

In practice:

- difficult to meet these 3 requirements simultaneously

4. Recognition and classification

4.1 Textual Features

Text representation

Linear text representation as a sequence of ...

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

Textual evidence for named entities appears at different levels:

- characters: **morphological clues** (word shape)
- words: **lexical clues** (syntactic and semantic information)
- word sequence: **contextual clues**

Interactive session

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

Capitalization

- widely used in Western character sets to mark a proper noun

BUT

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

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 , *-dóttir*
  - in North Africa, prefixes *Ben-* or *Aït-*
  - in Japanese, the suffix *-san*
- tokens originating from conventions or standards : *Inc.* in English, *S.A.* in French, *GmbH* in German

Presence of digits

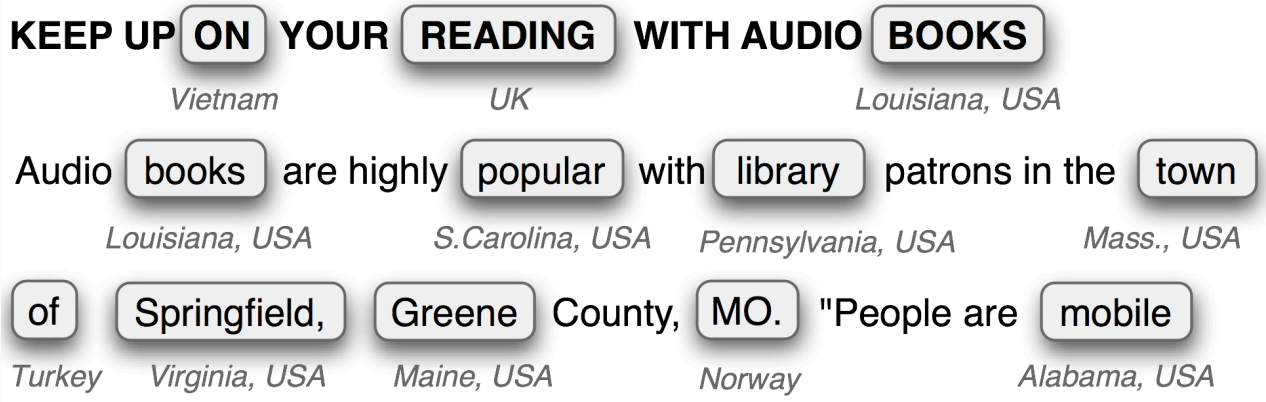
- dates, amounts or measures typically contain numbers (written in digits or letters)
- numbers can be written differently (*10 000*, *quatre-vingt-treize*, *cent dix-huit*) and/or with specific characters (*10,38*, *24/03*).
- mix of numeric and alphabetic characters (*100km*, *10h30*, etc.).
- abbreviations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

Lexical clues

Basics: identify textual words corresponding to lexicon entries (gazetteers, or other semantic resources).

- accurate if lexical entries are controlled (but maintenance is costly)
- NEs are an open class, new names appear constantly [McD96, Fri02]
- gazetteers/knowledge bases are organized according to NE types

Ambiguity: Gazetteers cannot be applied blindly...



Source: http://www.nltk.org/book/ch07.html

Contextual clues

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

$\rightarrow$ need additional clues nearby:

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

Helpful disambiguation hypothesis on document level

"One Sense per Discourse" [GCY92]

Importance of contextual clues

1. *He saw May on television.*
2. *His trip in May went well.*
3. *He bought a Renault Clio.*
4. *The Clio muse sings the past of humans and cities.*

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

Combining clues

Joint consideration of morphological, lexical and contextual features:

- **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.

Dealing with the **unreliability of clues**

- Clues are never 100% correct and reliable
- **Statistical models** for integrating them are needed

Complex JAPE rule for recognition of person names

```
Rule: PersonTitle
Priority: 35
/* allows Mr. Jones, Mr Fred Jones etc. */

(
 (TITLE)
(FIRSTNAME | FIRSTNAMEAMBIG | INITIALS2)*
 (PREFIX)?
 {Upper}
 ({Upper})?
 (PERSONENDING)?
)
:person -->
 {
FeatureMap features = Factory.newFeatureMap();
AnnotationSet personSet = bindings.get("person");

// get all Title annotations that have a gender feature
 HashSet fNames = new HashSet();
    fNames.add("gender");
    AnnotationSet personTitle = personSet.get("Title", fNames);

// if the gender feature exists
 if (personTitle != null && personTitle.size()>0)
{
  Annotation personAnn = personTitle.iterator().next();
  features.put("gender", personAnn.getFeatures().get("gender"));
}
else
{
```

Source:
https://gate.ac.uk/releases/gate-8.5.1/tao/splitch8.html

- JAPE formalism of GATE Information Extraction platform
- Regular expression combines morphological (uppercase), lexical (names, titles) and contextual (local grammar) clues
- Matched rule has action part for constructing the text annotation
- Sets a gender attribute if available in title or firstname

Typical feature sets for statistical NER

4.2.1 Spelling features

We extract the following features for a given word in addition to the lower case word features.

- whether start with a capital letter
- whether has all capital letters
- whether has all lower case letters
- whether has non initial capital letters
- whether mix with letters and digits
- whether has punctuation
- letter prefixes and suffixes (with window size of 2 to 5)
- whether has apostrophe end ('s)
- letters only, for example, I. B. M. to IBM
- non-letters only, for example, A. T. &T. to ..&
- word pattern feature, with capital letters, lower case letters, and digits mapped to 'A', 'a' and '0' respectively, for example, D56y-3 to A00a-0

Source: [HXY15]

Contextual clues typically include

- preceding, succeeding words
- parts of speech of these words

In Digital Humanities

- coverage and validity of lexical resources
- historical spellings, historical name variants and trigger words (e.g. occupation names)
- OCR errors

Textual features: take-home message

- Morphological, lexical or contextual features
- Possibility of combining and weighting the evidence by rules or statistical models
- these are the 'ingredients' of NE processing systems

4. Recognition and classification

4.2 Rule-based approaches

Developments in text technology

Rule-based systems

In the past, handwritten rule-based systems were the norm.

Machine learning based systems

Meanwhile, systems that learn inductively from annotated data are dominant in research. Methods based on neural networks (deep learning) are often particularly effective.

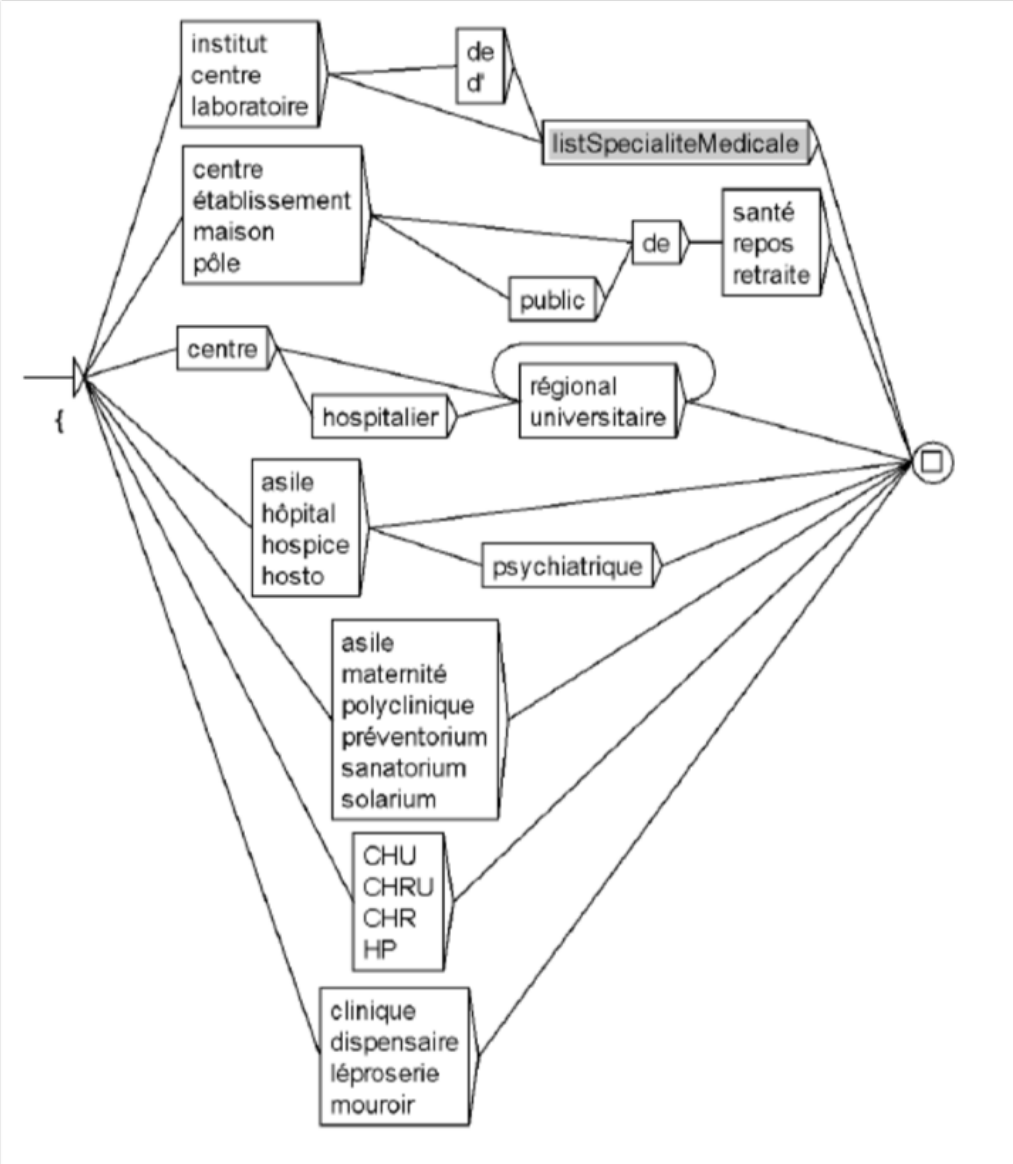
Unsupervised and adaptive systems

In order to reduce/avoid time-consuming annotation of data, procedures are sought which learn as much as possible unsupervised from data or adapt with data.

Rule-based approaches

- **Goal**: (possibly nested) markup (stand-in or stand-off) for mentions in text
- annotations can be complex feature structures with many attributes
- **Principle**: application-specific *rules* forming a *local grammar*
- **Development**: use of phrase matchers and/or transducers
- Several **frameworks**:
  - GATE
  - NooJ
  - Unitex
  - Spacy (supports modern neural and simple rule-based methods)

Automatons



- recognition of mentions by traversing a finite-state automaton from the start to the end (one-way streets)
- nodes specify textual material that needs to be seen
- move along connections while “reading” the text
- more to come in the hands-on session

4. Recognition and classification

4.3 Machine learning approaches

Rule-based vs. data-driven systems

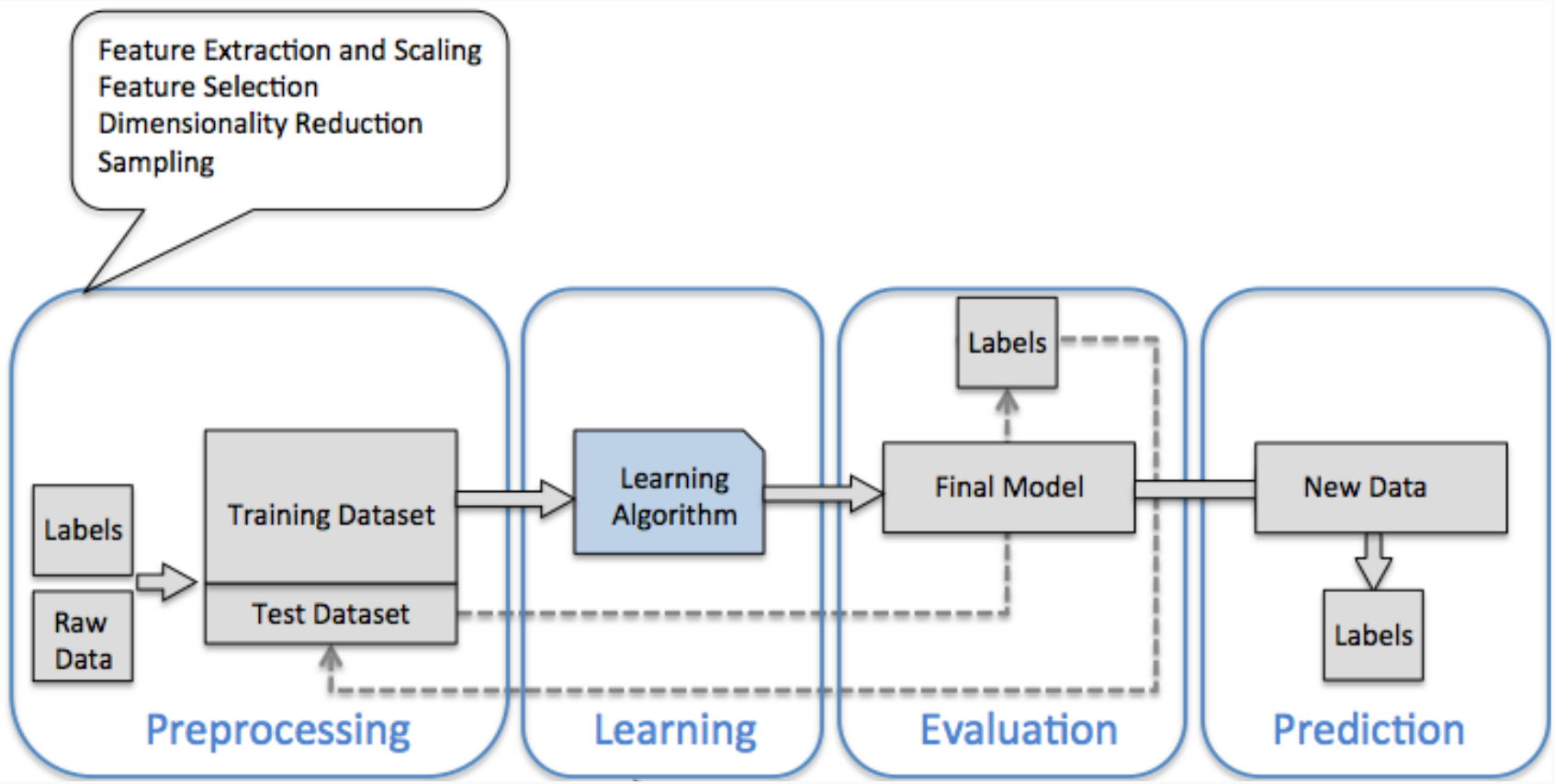
Rule-based systems

- Developer creates language-specific resources (gazetteers, automata, applications)
- Data is used for development and evaluation
- Developer is in control

Data-driven supervised systems

- Annotators create annotated corpora with target labels
- Tools are typically language-independent
- Statistical model learns (generalizes) from data
- Developer specifies features, statistical model, and learning algorithm
- Data is used for training, development and evaluation
- Annotated data is in control

Typical workflow for supervised machine learning



Source: [?, 11]

Feature engineering and learning algorithms are often complex in NLP!

What is (probabilistic) classification?

Hard classification: $y = f(\mathbf{x})$

$$\mathrm{CLASS} = f(\mathrm{FEATURES})$$

Class Nominal category

Features Nominal, ordinal or numeric data

f classification function (*predictor function*)

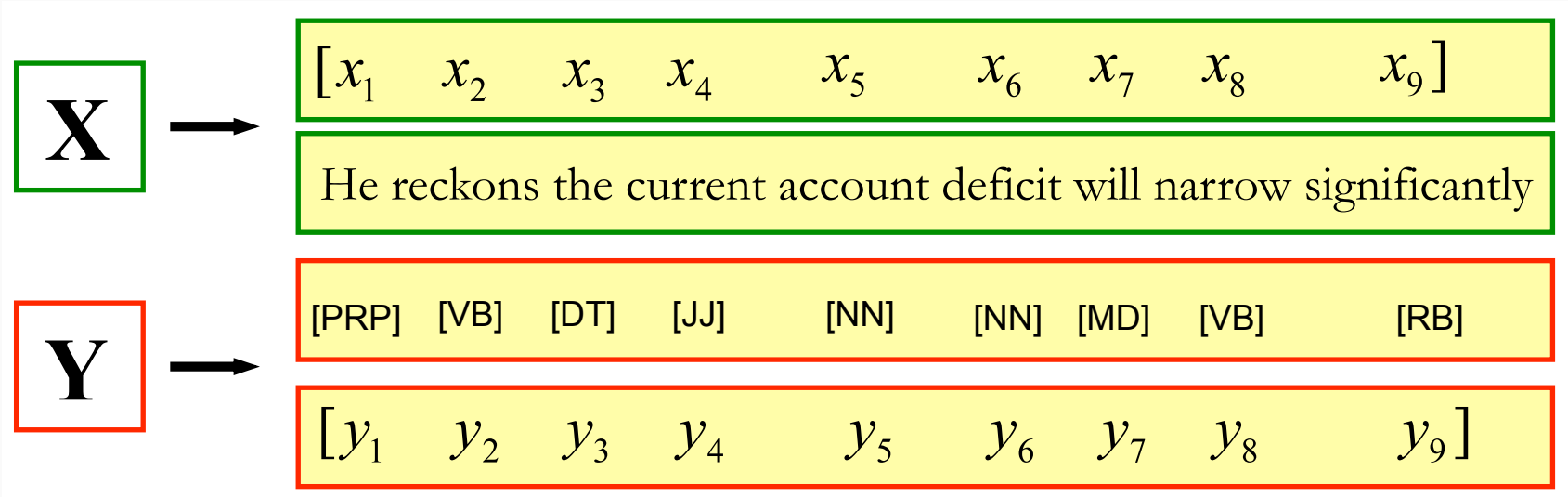
Class	**Features**
Outcome	Evidence
Output Variable	Input Variable
Dependent Variable	Independent Variable

Probabilistic classification: $P(\mathrm{CLASS}|\mathrm{FEATURES})$

Most probable class: $\hat{y} = \arg\max_y P(Y = y|\mathbf{x})$

Classical NLP example for sequence classification

But texts are not independent bag-of-words...



Wanted: Most likely POS tag sequence given the word sequence

$$y^*_{1..n} = \arg\max_{y_{1..n}} p(y_{1..n}|x_{1..n})$$

Dependencies between neighbouring words are characteristic for texts!

NER problem reformulation

In [DATE 2000], [PER Monica Belucci] presented [MISC Under Suspicion] by [PER Stephen Hopkins].

Subproblems of NER

- Segment text in name mentions
- Classify segmented mentions into NER categories

Can we reformulate NER as a sequence classification ?

And solve both subproblems in one go?

NER tagging as sequence labeling: IOB coding

X	Y
Token	NER tag
In	O
2000	B-DATE
,	O
Monica	B-PER
Belucci	I-PER
presented	O
Under	B-MISC
Suspicion	I-MISC
by	O
Stephen	B-PER
Hopkins	I-PER
.	O

IOB schema (BIO2)

B-C Begin of a name of class C

I-C Inside a name of class C

O Outside of a name

Learn from annotated data

Any ML technique for sequence labeling can be used for NER!

Strong dependencies between labels

I-CLASS tags need preceding tags of a certain type!

Feature functions in Multinomial logistic regression (MaxEnt)

Feature function for categorical features

$$f_i(y, \mathbf{x}) = \begin{cases} 1 & \text{if } y = \text{BPER} \wedge x_{\text{FirstCharIsUpper}} = \mathit{True} \\ 0 & \text{otherwise} \end{cases}$$

Exponential classification model

$$P(y|\mathbf{x}, \lambda) = \frac{1}{Z_{\mathbf{x}}} \exp\left(\sum_{i=1}^{K} \lambda_i f_i(y, \mathbf{x})\right)$$

- λ_i Weight of feature f_i
- K Total number of features
- Z_x Normalization: $Z_x = \sum_y \exp\left(\sum_i \lambda_i f_i(y, \mathbf{x})\right)$

Goal of learning algorithm

Optimize weights in order to make training data as probable as possible!

Feature functions in Conditional Random Fields (CRF)

In an exponential linear-chain conditional sequence classification model, each feature function $f(y_{i-1}, y_i, \mathbf{x}, i)$ takes

- an input sequence **x**
- the position i of a word in **x**
- the label y_i of the current word
- the label y_{i-1} of the previous word

Evidence in input sequence x

Can be more than just word form: lemma, part-of-speech, syntactic dependency label, semantic class etc.

See excellent blog on practical ML techniques and NER![1]

[1] https://www.depends-on-the-definition.com/named-entity-recognition-conditional-random-fields-python/

Exponential Conditional Sequence Model (CRF)

Label sequence modelled as a normalized product of feature functions:

$$P(\mathbf{y} \mid \mathbf{x}, \boldsymbol{\lambda}) = \frac{1}{Z(\mathbf{x})} \exp \sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i-1}, y_i, \mathbf{x}, i)$$

$$Z(\mathbf{x}) = \sum_{\mathbf{y} \in Y} \sum_{i=1}^{n} \sum_{j} \lambda_j f_j(y_{i-1}, y_i, \mathbf{x}, i)$$

The model is log-linear on the Feature Functions

Source: [?]

Every feature function f_i is weighted (parameterized) by its weight λ_i!

Probabilistic sequence classification

Select the most probable label sequence $\mathbf{y}^*$!

Sequence classification: take-home message

- NER can be cast as a sequence labeling problem
- CRFs consider dependencies between subsequent labels
- Classical CRFs use features functions to encode symbolic features numerically

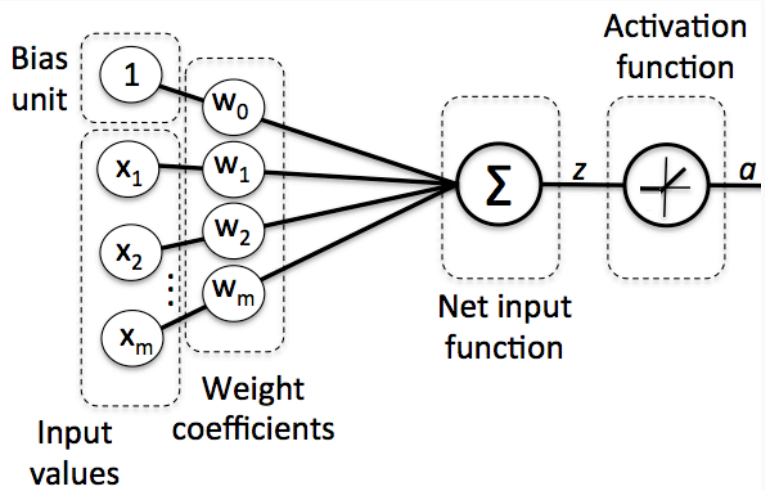
4. Recognition and classification

4.4 Neural approaches

Components of a single neuron

Computations of a Neuron

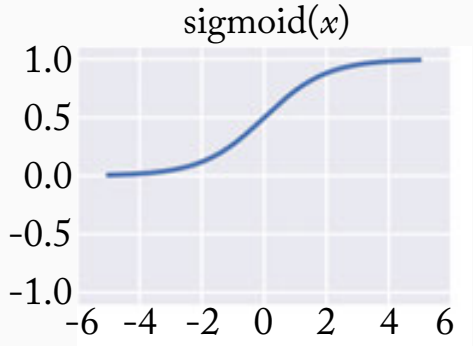
- Scalar numeric inputs
- Weights for inputs
- (Nonlinear) activation function

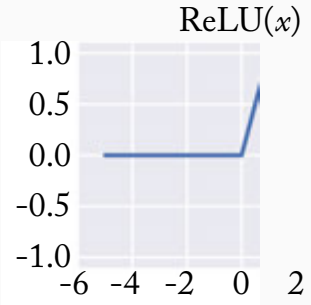


Weighted sum of inputs...

Activation functions

There are many...



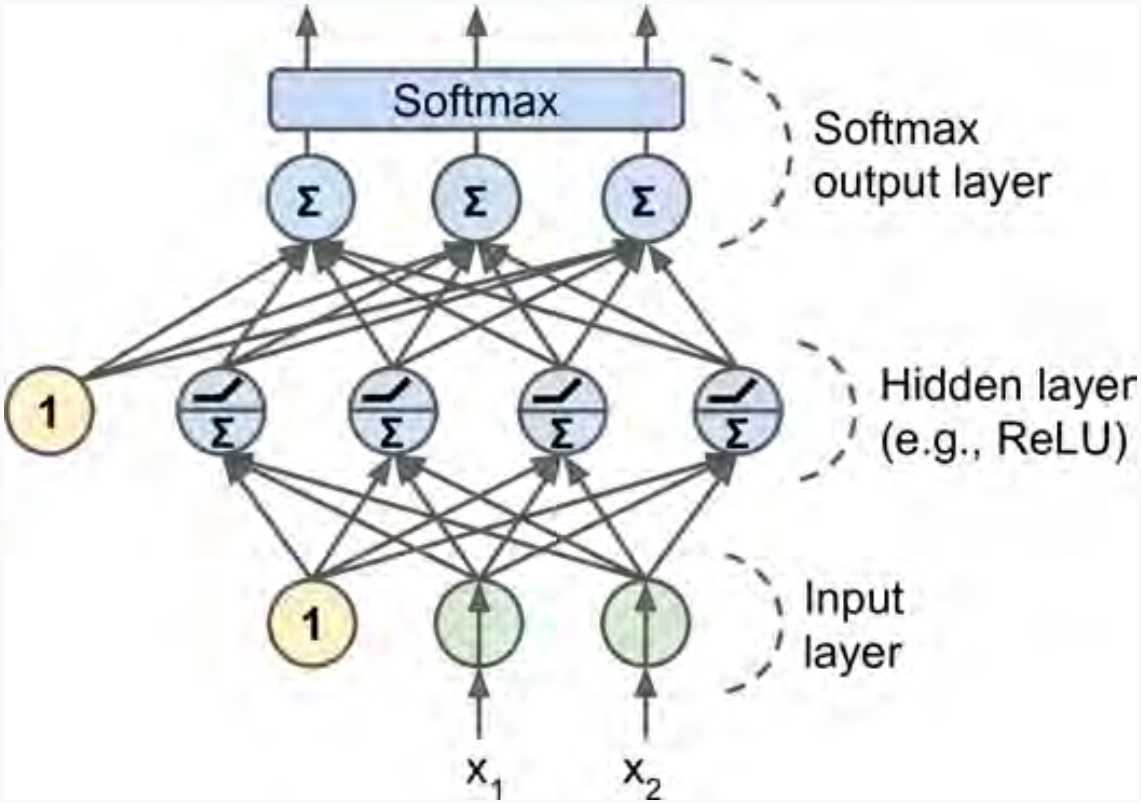


...and scaling the output

$$a = sigmoid(1 * w_0 + x_1 * w_{1*...} + x_n * w_n)$$

Feedforward networks for probabilistic classification

Multiple neurons next to each other (layers) and after each other (deepness).



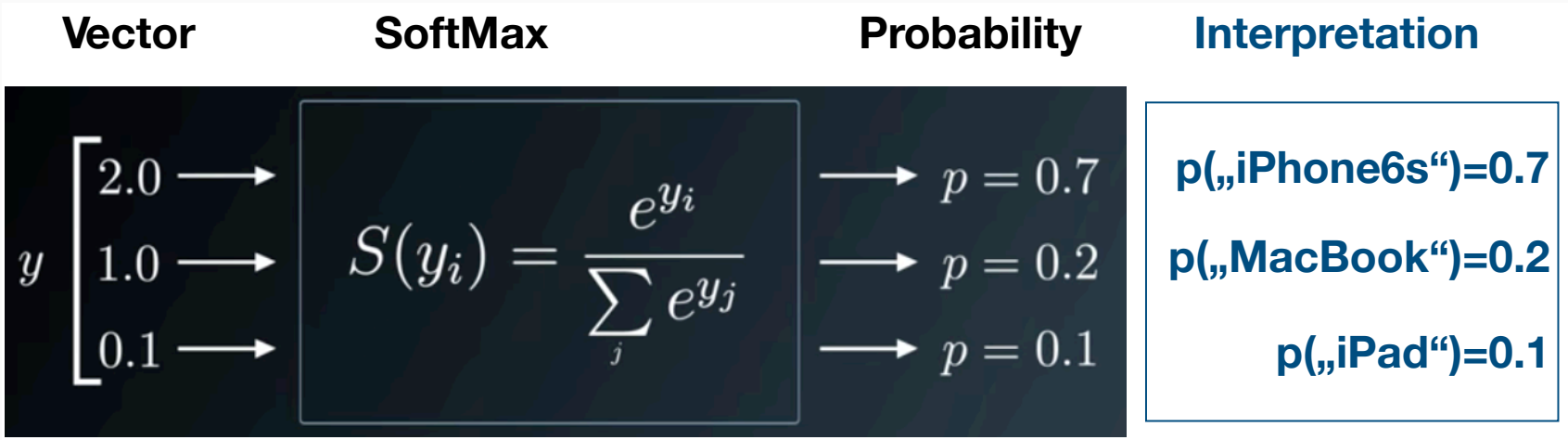
Source: [Gé17]

Each output neuron emits the probability of a class!

SoftMax: Turning vectors into probabilities

SoftMax: hello again, exponential model...

Every vector with *n* numbers can be normalized into a probability distribution over *n* classes.



Remaining problem: how can words be turned into numeric input?

4. Recognition and classification

4.5 Word Embeddings: Dense word representations

Corpus-based distributionalism [Sah08]

An old idea that profits from big data

- "You shall know a word by the company it keeps!" (J. R. Firth (1957))
- "words with similar meanings will occur with similar neighbors if enough text material is available"

(Schütze & Pedersen, 1995)

A small corpus

- I like deep learning.
- I like NLP.
- I enjoy flying.

Idea

Similar words have similar rows.

Bigram statistics

counts	I	like	enjoy	deep	learning	NLP	flying	.
I	0	2	1	0	0	0	0	0
like	2	0	0	1	0	1	0	0
enjoy	1	0	0	0	0	0	1	0
deep	0	1	0	0	1	0	0	0
learning	0	0	0	1	0	0	0	1
NLP	0	1	0	0	0	0	0	1
flying	0	0	1	0	0	0	0	1
.	0	0	0	0	1	1	1	0

Source: [Soc16, 9]

Many ways of distributional modeling

tokenization
annotation
tagging
parsing
feature selection

⋮ cluster texts by date/author/discourse context/...

↓ ↙

Matrix type		Weighting		Dimensionality reduction		Vector comparison
word × document		probabilities		LSA		Euclidean
word × word		length normalization		PLSA		Cosine
word × search proximity	×	TF-IDF	×	LDA	×	Dice
adj. × modified noun		PMI		PCA		Jaccard
word × dependency rel.		Positive PMI		IS		KL
verb × arguments		PPMI with discounting		DCA		KL with skew
⋮		⋮		⋮		⋮

Source: [Pot13]

Categorical word representations: one-hot encoding

The standard word representation

The vast majority of rule-based **and** statistical NLP work regards words as atomic symbols: hotel, conference, walk

In vector space terms, this is a vector with one 1 and a lot of zeroes

[0 0 0 0 0 0 0 0 0 0 1 0 0 0 0]

Dimensionality: 20K (speech) – 50K (PTB) – 500K (big vocab) – 13M (Google 1T)

We call this a "one-hot" representation. Its problem:

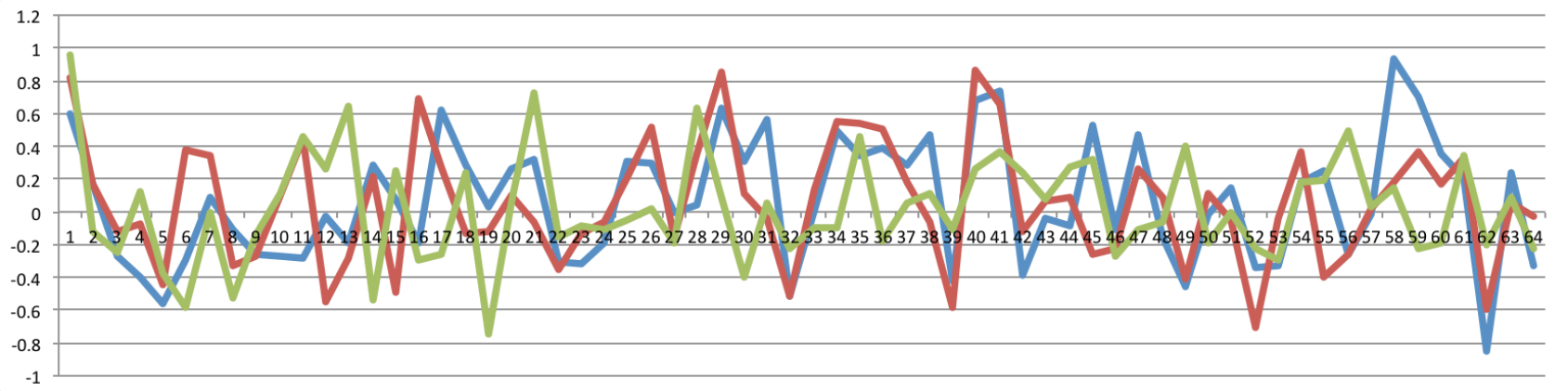
motel [0 0 0 0 0 0 0 0 0 0 1 0 0 0 0] AND
hotel [0 0 0 0 0 0 0 1 0 0 0 0 0 0 0] = 0

Source: [SM13, 35]

Question: How similar▲ are they distributionally?

Words as dense vectors of 64 real numbers

Which line represents which word? (“war”, “peace”, “tomato”)



Word Embeddings

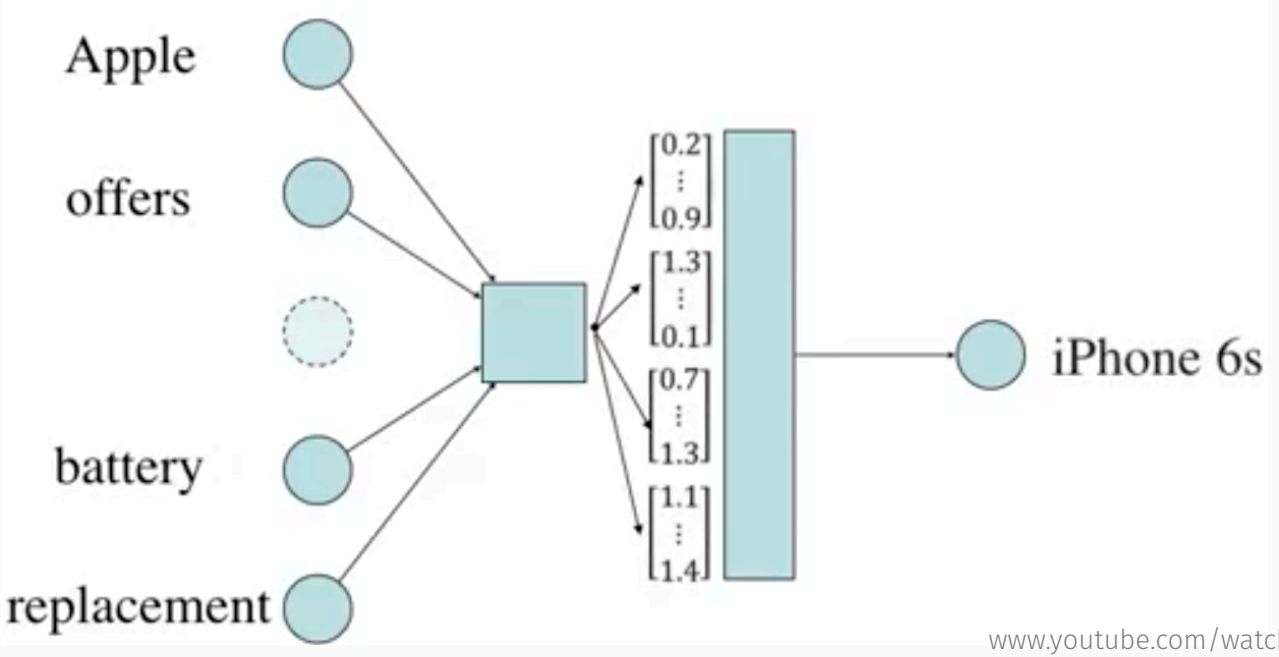
Continuous, numeric, dense word representations learned from raw text.

Similar vectors mean similar words (cosine vector distance)

Embeddings are a *perfect input* for numeric ML methods.

CBOW: Word prediction task

Given a context of 4 surrounding words, what is the probability of seeing the center word *w*? [?]

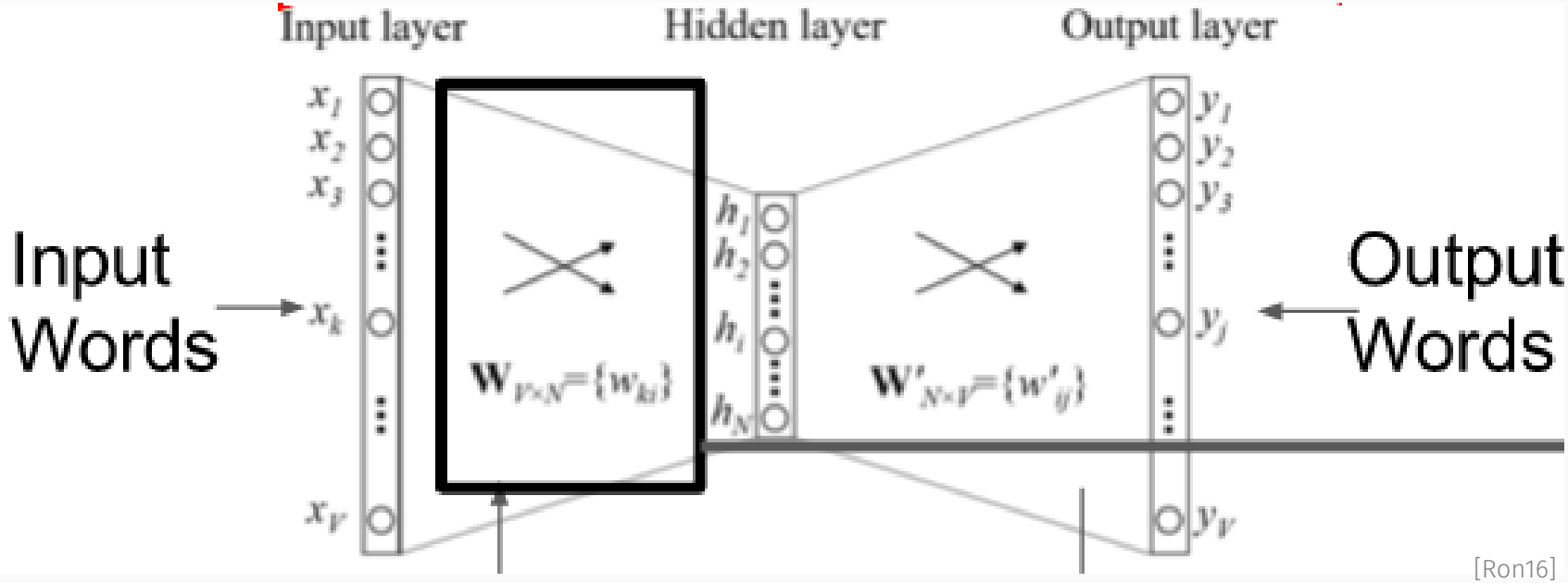


www.youtube.com/watch?v=aZarigloqXc

More theoretical insights on the exact distributional model of `word2vec` in [LG14]

Word embeddings as dense hidden layers for one-hot-vectors

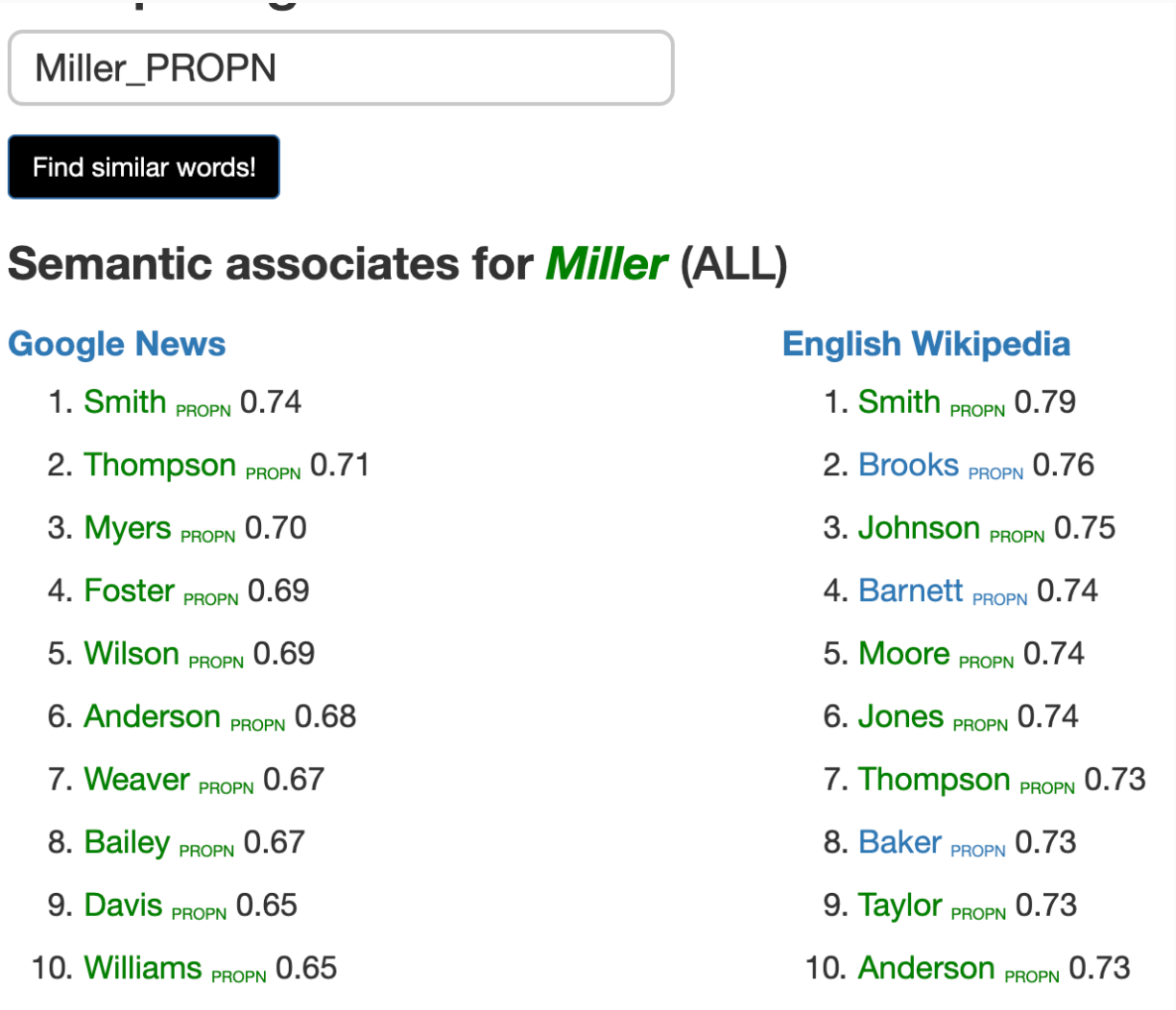
Neural Network learns to estimate the probability



[Ron16]

- Hidden layer has dimensionality of embeddings
- Input and output layer have dimensionality of one-hot-encoded vocabular

Nearest neighbors in dense vector space



Surnames are close in semantic space

Helps to generalize the annotated training material in a data-driven way

Given enough data, “gazetteer-style” information emerges

Embeddings of ambiguous words

- Classical embeddings operate at the level of word types (sometimes disambiguated by part-of-speech)
- Representation of ambiguous words are “representative” mixtures of the occurrences of the different senses in the corpus.
- parts of multiword expressions are often close to each other

Semantic associates for *Washington* (ALL)

British National Corpus	English Wikipedia
1. Us PROPN 0.62	1. D.C. PROPN 0.88
2. Bush PROPN 0.58	2. Washington D.C. PROPN 0.79
3. United PROPN 0.57	3. Seattle PROPN 0.75
4. Smithsonian PROPN 0.57	4. Washington , DC PROPN 0.74
5. Pentagon PROPN 0.57	5. Washington DC PROPN 0.72
6. States PROPN 0.56	6. Tacoma PROPN 0.70
7. Reagan PROPN 0.55	7. Spokane PROPN 0.68

Word representations: take-home message

Symbolic representations of words lead to high-dimensional and sparse vectors that are not suitable for expressing similarities.

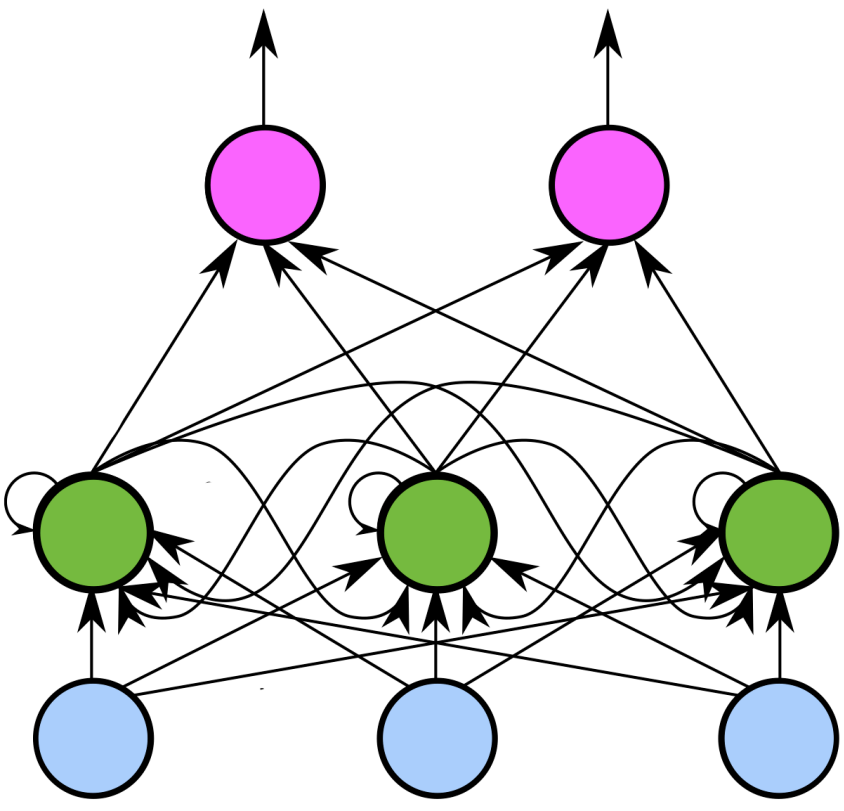
Continuous, low-dimensional dense representations result in more suitable semantic spaces for many NLP methods.

Classical word embeddings mix different senses of ambiguous words

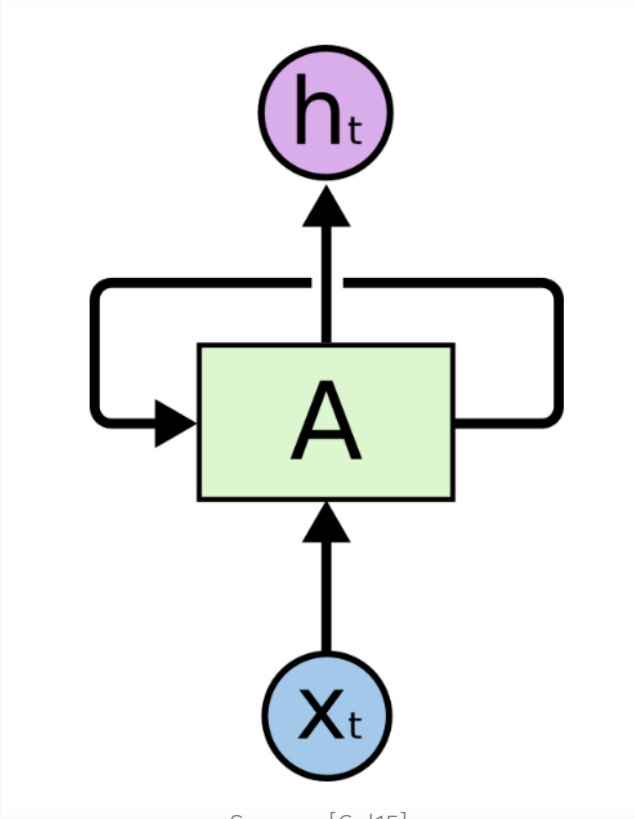
4. Recognition and classification

4.6 Neural sequence labeling

Elman Recurrent Neural Networks (RNN): feedback loops



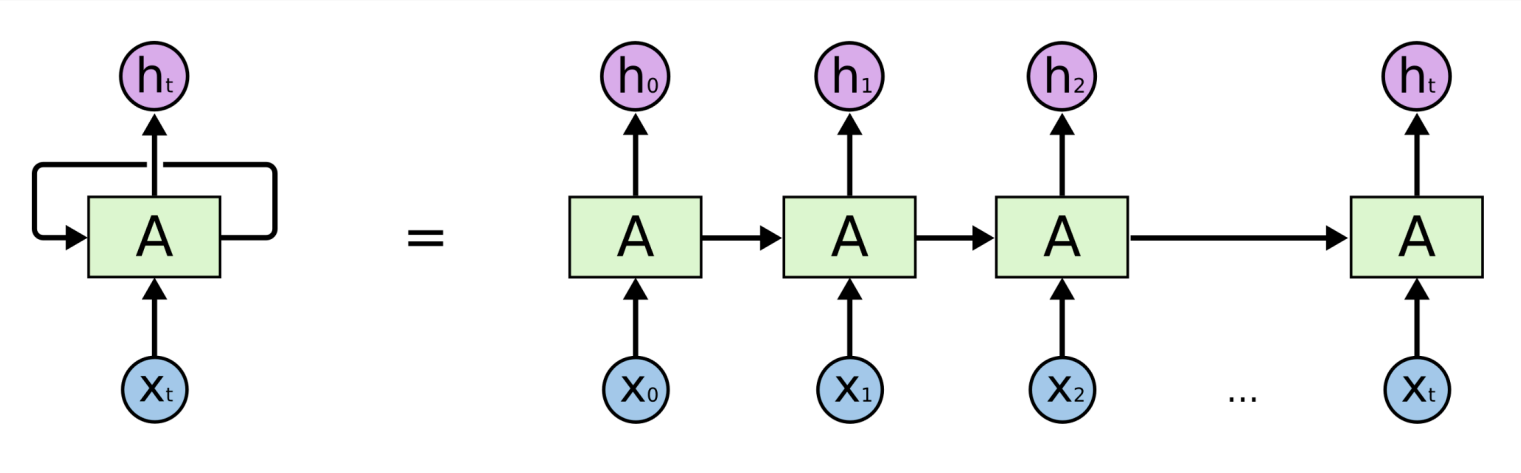
All neurons of hidden layer have a recurrent connection!
Information can persist over time!



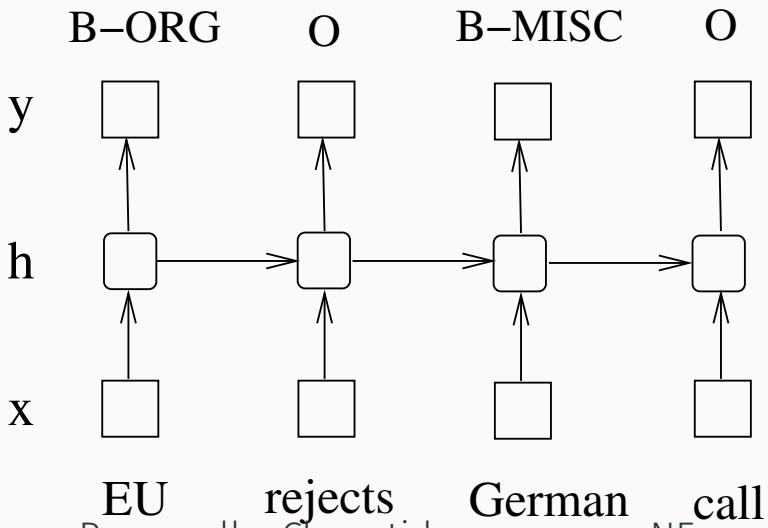
Source: [Col15]

Schematic drawing of feedback loop

Unrolled RNNs: processing sequences step by step



A NER tagging perspective [HXY15]



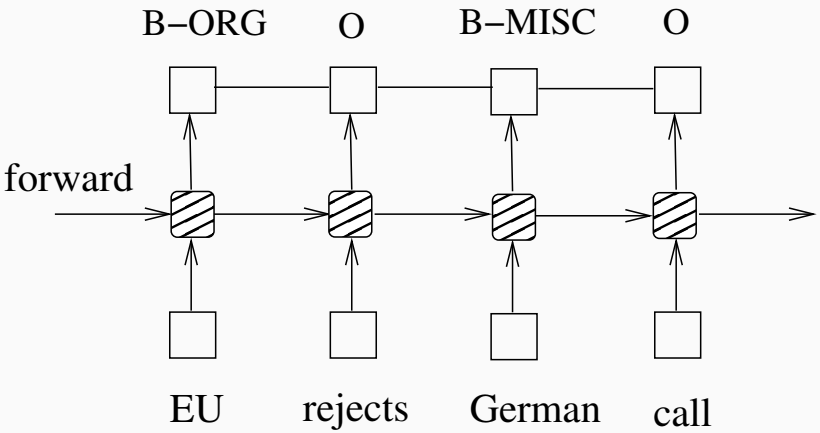
- At each processing step *t*, emit the most probable NER tag!
- Analog to separate MaxEnt classification of each word

Source: [HXY15]

Long Short-Term Memory Networks (LSTM)

- Powerful and complex variant of recurrent NNs (RNN)
- Gated RNN architectures can learn when to consider past input and when to forget it
- Deal well with local dependencies between inputs.
- Learn non-local dependencies between inputs much better than normal Elman RNNs.
- Have proven to be extremely performant in various NLP tasks.
- Exist in different mathematical variants (GRU) [GSK+16]

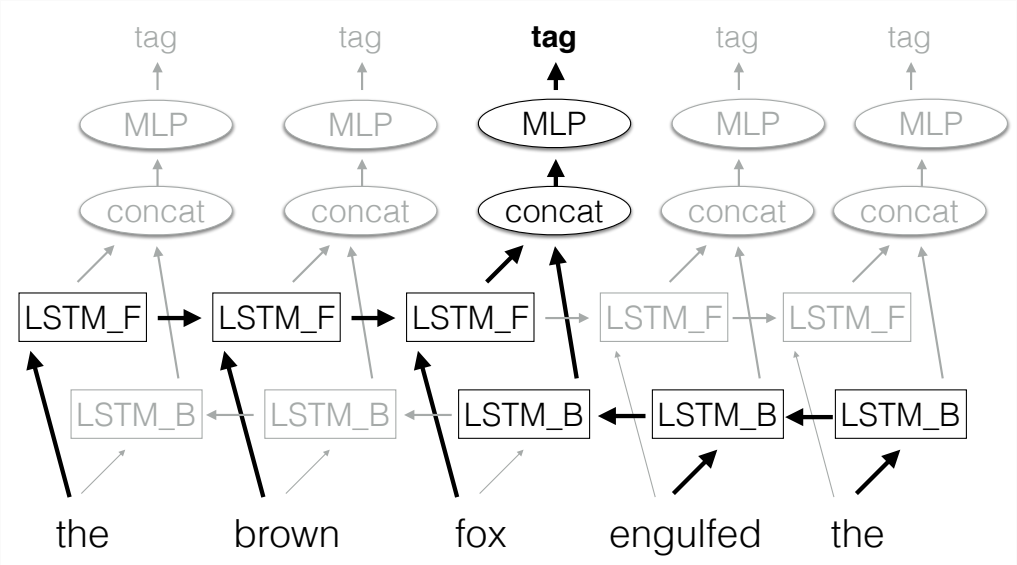
LSTMs with CRF-style output layer



- Network learns to output most probable NER tag sequence
- Analog to CRF objective!

Basic ideas of a BiLSTM tagger [HXY15]

- Learn task-specific LSTM embeddings over the sequence of tokens from left to right and right to left ...
- with vector concatenation of the two representations for each token.
- BiLSTM representation of each token can be used by Feedforward NN (also Multilayer Perceptron, MLP) to predict each tag independently.

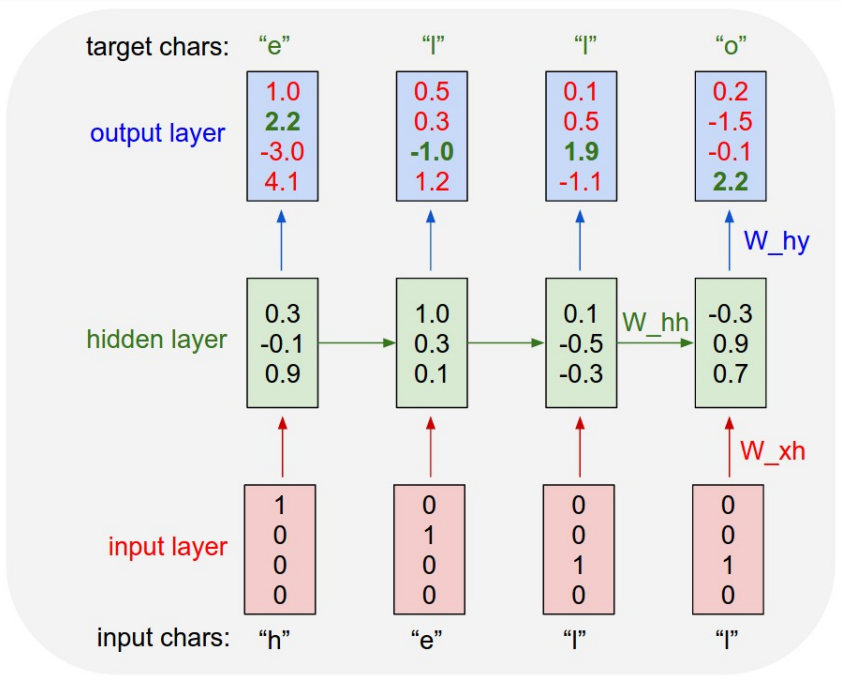


Character-Level language models

Shannon game [Sha51]

Guess the next characte?

Also a case of probabilistic sequence classification:
Learn to predict the next character!



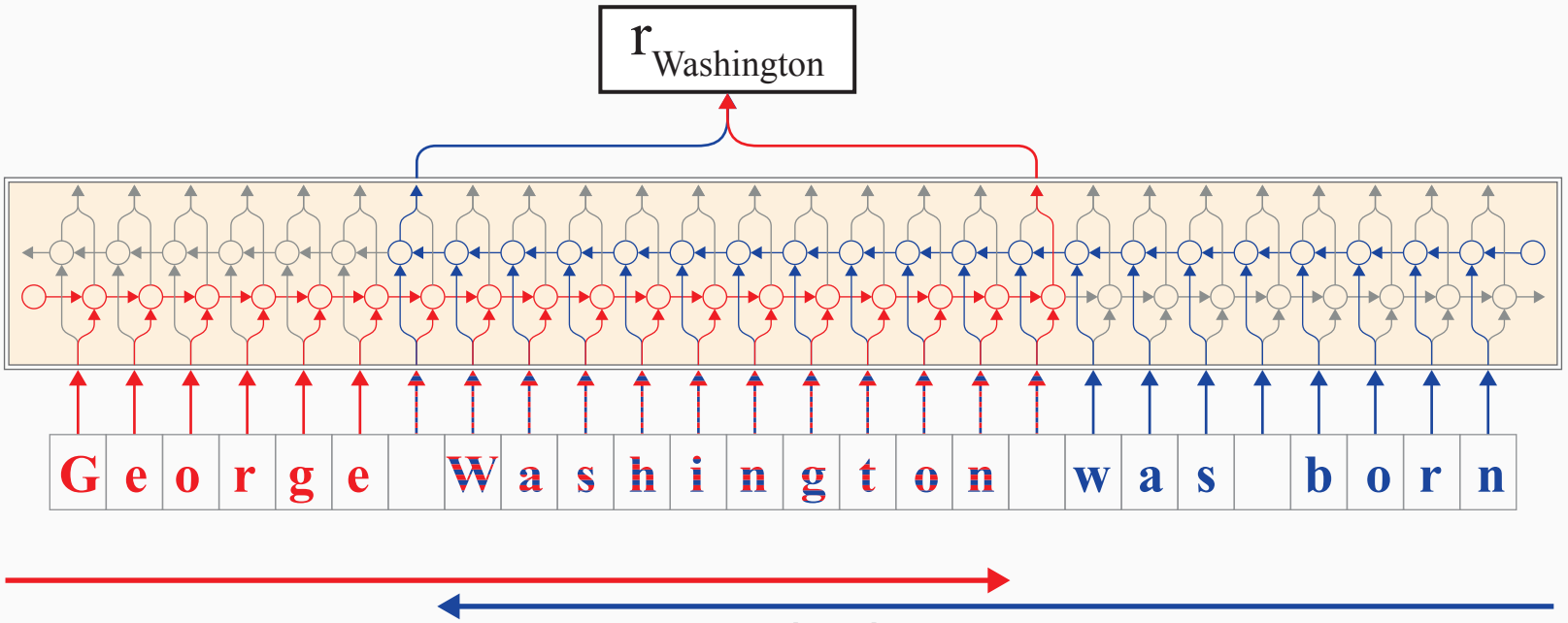
Source: Famous blog [Kar15]

- Alphabet has only 4 characters: h,e,l,o
- Green number in output layer indicates correct character

Generated text when trained on Wikipedia

Copyright was the succession of independence in the slop of Syrian influence that was a famous German movement based on ...

Contextualized string embeddings: flair embeddings



Source: [ABV18]

- Use forward and backward LSTM character language model
- Concatenation of hidden states reached after last character of a word was read
- Context integrates naturally into word representation
- Robust for noisy texts

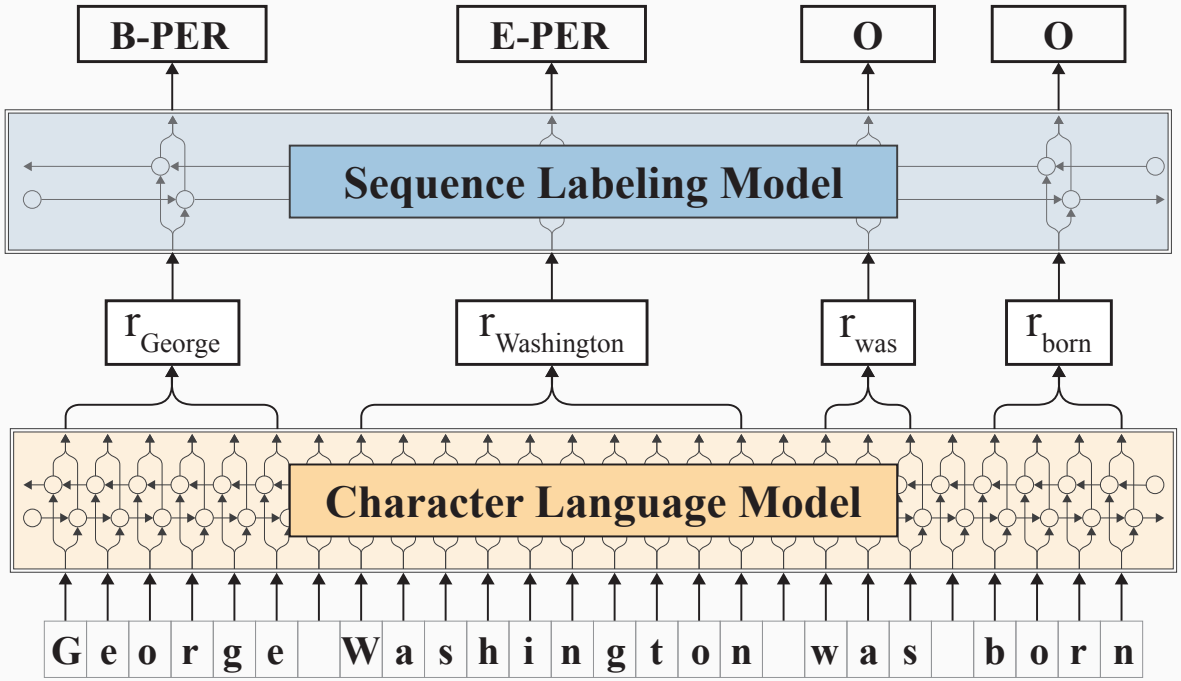
Contextualized string embeddings: context matters

word	context	selected nearest neighbors
Washington	(a) ***Washington*** *to curb support for [..]*	(1) ***Washington*** *would also take [..] action [..]* (2) ***Russia*** *to clamp down on barter deals [..]* (3) ***Brazil*** *to use hovercrafts for [..]*
Washington	(b) *[..] Anthony* ***Washington*** *(U.S.) [..]*	(1) *[..] Carla* ***Sacramento*** *(Portugal) [..]* (2) *[..] Charles* ***Austin*** *(U.S.) [..]* (3) *[..] Steve* ***Backley*** *(Britain) [..]*
Washington	(c) *[..] flown to* ***Washington*** *for [..]*	(1) *[..] while visiting* ***Washington*** *to [..]* (2) *[..] journey to New York City and* ***Washington*** *[..]* (14) *[..] lives in* ***Chicago*** *[..]*
Washington	(d) *[..] when* ***Washington*** *came charging back [..]*	(1) *[..] point for victory when* ***Washington*** *found [..]* (4) *[..] before* ***England*** *struck back with [..]* (6) *[..] before* ***Ethiopia*** *won the spot kick decider [..]*
Washington	(e) *[..] said* ***Washington*** *[..]*	(1) *[..] subdue the never-say-die* ***Washington*** *[..]* (4) *[..] a private school in* ***Washington*** *[..]* (9) *[..] said* ***Florida*** *manager John Boles [..]*

Table 4: Examples of the word "Washington" in different contexts in the CoNLL03 data set, and nearest neighbors using cosine distance over our proposed embeddings. Since our approach produces different embeddings based on context, we retrieve different nearest neighbors for each mention of the same word.

Source: [ABV18]

NER tagging architecture of flair



Source: [ABV18]

- Contextualized representation of each word enters NER sequence labeling model
- NER model is also a bi-direction LSTM
- The decision at every word integrates evidence from the full sequence

Stacking of different word embeddings

Class	Type	Pretrained?
`WordEmbeddings`	classic word embeddings (Pennington et al., 2014)	yes
`CharacterEmbeddings`	character features (Lample et al., 2016)	no
`BytePairEmbeddings`	byte-pair embeddings (Heinzerling and Strube, 2018)	yes
`FlairEmbeddings`	character-level LM embeddings (Akbik et al., 2018)	yes
`PooledFlairEmbeddings`	pooled version of FLAIR embeddings (Akbik et al., 2019b)	yes
`ELMoEmbeddings`	word-level LM embeddings (Peters et al., 2018a)	yes
`ELMoTransformerEmbeddings`	word-level transformer LM embeddings (Peters et al., 2018b)	yes
`BertEmbeddings`	byte-pair masked LM embeddings (Devlin et al., 2018)	yes
`DocumentPoolEmbeddings`	document embeddings from pooled word embeddings (Joulin et al., 2017)	yes
`DocumentLSTMEmbeddings`	document embeddings from LSTM over word embeddings	no

Table 1: Summary of word and document embeddings currently supported by FLAIR. Note that some embedding types are not pre-trained; these embeddings are automatically trained or fine-tuned when training a model for a downstream task.

flair supports a whole zoo of powerful new embeddings variants.

Results of flair architecture on sequence labeling tasks

Approach	NER-English F1-score	NER-German F1-score	Chunking F1-score	POS Accuracy
proposed				
PROPOSED	91.97±0.04	85.78 ± 0.18	96.68±0.03	97.73±0.02
PROPOSED$_{+\text{WORD}}$	93.07±0.10	88.20 ± 0.21	96.70±0.04	97.82±0.02
PROPOSED$_{+\text{CHAR}}$	91.92±0.03	85.88 ± 0.20	**96.72**±0.05	97.8±0.01
PROPOSED$_{+\text{WORD+CHAR}}$	**93.09**±0.12	**88.32** ± 0.20	96.71±0.07	97.76±0.01
PROPOSED$_{+\text{ALL}}$	92.72±0.09	n/a	96.65±0.05	**97.85**±0.01
baselines				
HUANG	88.54±0.08	82.32 ± 0.35	95.4±0.08	96.94±0.02
LAMPLE	89.3±0.23	83.78 ± 0.39	95.34±0.06	97.02±0.03
PETERS	92.34±0.09	n/a	96.69±0.05	97.81± 0.02
best published				
	92.22±0.10	78.76	96.37±0.05	97.64
	(Peters et al., 2018)	(Lample et al., 2016)	(Peters et al., 2017)	(Choi, 2016)
	91.93±0.19	77.20	95.96±0.08	97.55
	(Peters et al., 2017)	(Seyler et al., 2017)	(Liu et al., 2017)	(Ma and Hovy, 2016)
	91.71±0.10	76.22	95.77	97.53±0.03
	(Liu et al., 2017)	(Gillick et al., 2015)	(Hashimoto et al., 2016)	(Liu et al., 2017)
	91.21	75.72	95.56	97.30
	(Ma and Hovy, 2016)	(Qi et al., 2009)	Søgaard et al. (2016)	(Lample et al., 2016)

Table 2: Summary of evaluation results or all proposed setups and baselines. We also list the best published scores for each task for reference. We significantly outperform all previous works on NER, and slightly outperform the previous state-of-the-art in PoS tagging and chunking.

Neural end-to-end sequence labeling: take-home message

Character and word embeddings trained on large text corpora contain a lot of morphological and lexical information

End-to-end optimization of NNs enable an effective representation learning; Bi-LSTMs have full access to contextual cues.

Explicit feature engineering can be reduced to a minimum

Relatively small amounts of task-specific annotation data give good performance

5. Linking

Where are we

- We know how to **recognize** and **classify** textual segments
- What remains to be done: **establish the link** between mentions and the objects to which they refer

$\longrightarrow$ Objective: disambiguation, resolution, linking

From mention to referents

- **Categorizing does not mean we know everything**:
  G. Bush and *F. Mitterrand* are PERSON
  But which refers to the *43th US president*?

- **Homonymy** (one name, several referents):
  F. Mitterrand is a PERSON
  But *François Mitterrand* or *Frédéric Mitterrand* ?

  Bush is a PERSON
  But *G. W. Bush* or *G. Bush* ?

- **Name variants** (one referent, several names):
  Do *Jean-Claudem Junckerem*, *Juncker*, *Jean-Cluade Juncker* and *the president of the Europena Commission* refer to the same entity?

Tasks

- **Co-reference resolution**:
  within a document, identify that *Frédéric Mitterrand*, *Mitterrand*, *FM* have the same referent.
- **Mention clustering**:
  within a collection of documents, identify that *Frédéric Mitterrand*, *Mitterrand*, *FM* have the same referent .
- **Entity linking**:
  within a collection of documents, identify entity mentions and **link** them to a referent in a knowledge base, or NIL if not present.

Formalisation

Given:

- the set of mentions in a corpus
- the set of entities in a knowledge base

Entity linking tries to match the two sets.

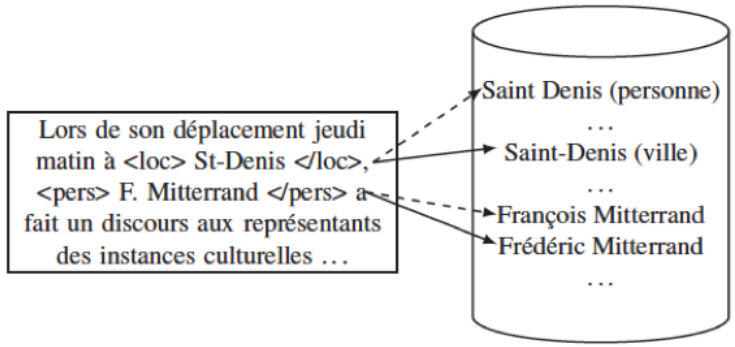


Figure 5.1. *Mentions du texte et références de la base de connaissances*

Steps

1. Mention detection in texts;
2. Candidate selection in the KB;
3. Linking.

Mention detection

1. Usage of NERC systems

2. Or simple look-up of very large collection of mentions

Candidate selection

- Confronting each mention to all referents is expensive and inefficient.
  $\rightarrow$ for each mention, candidate referents are selected.
- **Approach**:
  - on text side: consideration of the tokens composing mentions;
  - on KB side: consideration of name variants of existing referents.

  $\rightarrow$ for a given mention, selection of candidates whose name variants contain a token of the mention.
- **Example**: the mention *F. Mitterrand* has as candidates all person referents whose first name start with *F* and whose last name is *Mitterrand*

Candidate selection

Relaxed constraint on surface forms:

- case insensitive;
- removal of stop words, of parenthetical elements ;
- automatic generation of acronyms from surface forms ;
- etc.

$\rightarrow$ multiplying KB referent variants in order not to miss candidates.

Linking

Objective:
associate each mention with the most likely candidate (or NIL)

Approach:
consideration of clues on both sides (mention and candidate) and computation of a distance

Clues

- **Mentions**:
  - tokens
  - local context
  - global context
- **Candidates**:
  - textual description of the referent (title, synonyms, summary, article)
  - other properties (e.g. infoboxes)
  - associated entities and concepts

A lot of information is available, one needs to choose the most relevant.

Computation of similarity

- **Textual clues**
  $\rightarrow$ cosine distance
  (but does not work in all cases, e.g. similar entities)
- **Structural clues**:
  $\rightarrow$ selection of the most popular referent according to a criterion.
  - number of links pointing to the page of the referent
  - for a city, population size

Linking: take-home message

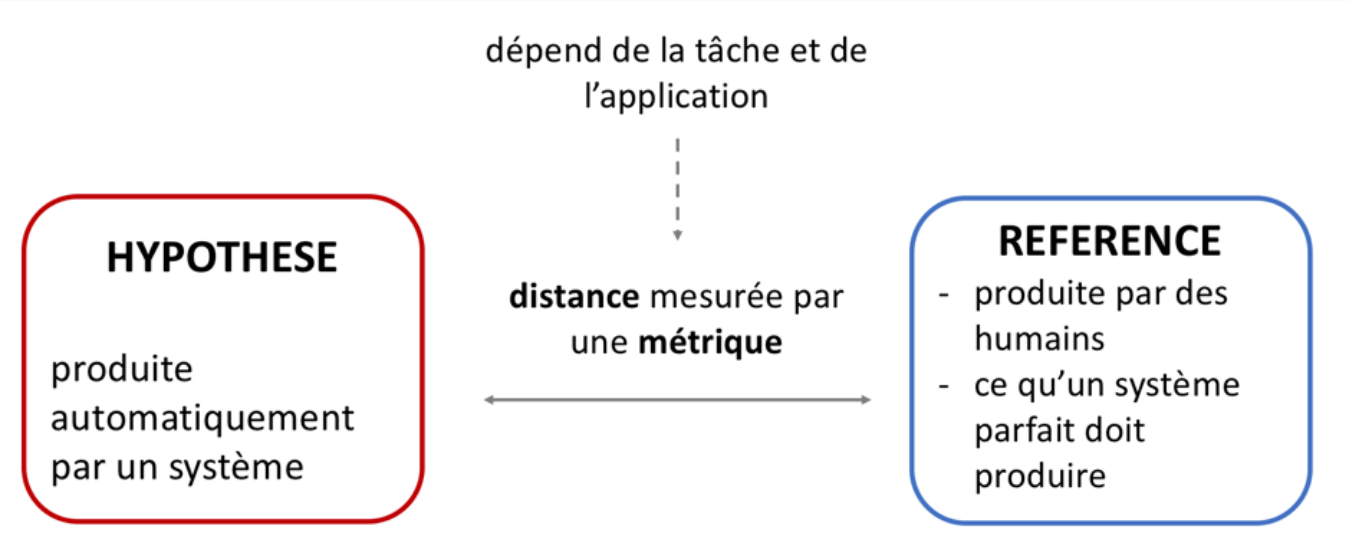
- Links between mentions and referents in a knowledge base
- Expansion of the task from ca. 2007 on
- Relative good performances for English
- Wikipedia is the main, if not the only, knowledge base
- Yet another building block to better understand texts

6. Evaluation

Evaluation

- First formalization of the evaluation procedure: MUC3 [Sun91]
- Motivation:
  - to have stable and effective elements of comparison between system hypothesis and references
  - comparison of systems
  - evaluation reproducibility beyond the evaluation campaign

Evaluation protocol



Objective: measure if the system find the 'good answers'

What is a 'good answer'?

- automatic translation and summarization: multiples good answers
- NE: we can assume a single "good answer"

Advantages

- **Transparency**: 'judgement' rules known to all;
- **Cost**: reduced cost compared to manual evaluation;
- **Reproducibility**: possibility of scientific results comparison.

What you need to evaluate

1. Metric(s) to measure the distance between the hypothesis and the reference
2. Alignment and projection algorithm between hypothesis and reference

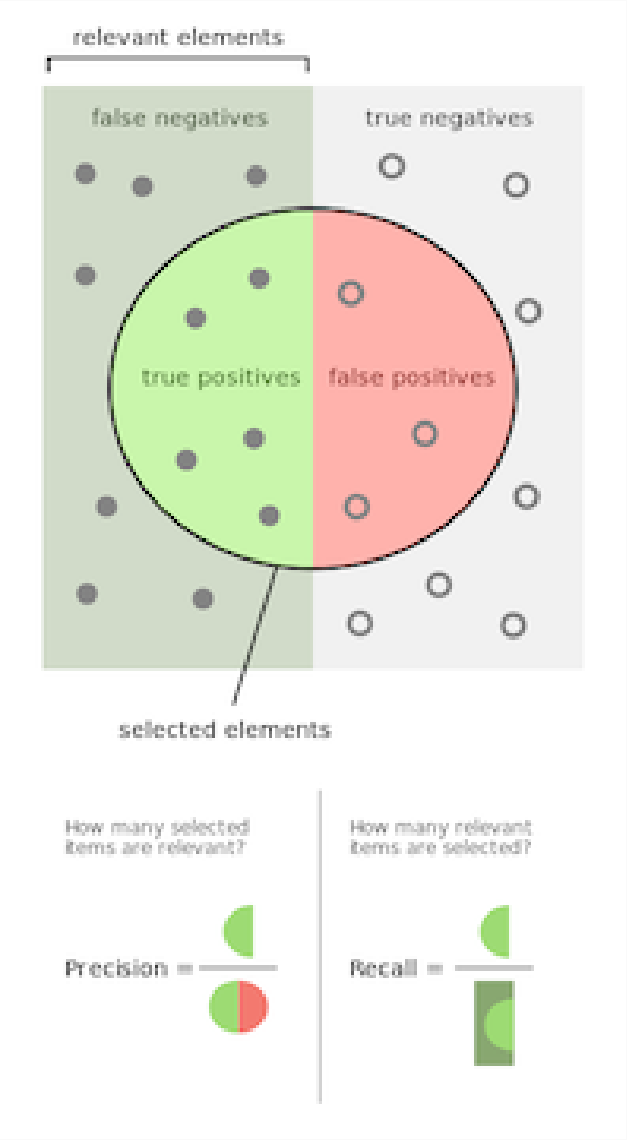
Classic measures

Precision

Ratio between the number of good answers and all the answers given by a system.

$$P = \frac{tp}{tp + fp} \tag{1}$$

- *tp*: **True positives**: number of **corrects** items in hypothesis;
- *fp*: **False positive**s: number of **insertions** (false alarm)



Classic measures

Recall

Ratio between the number of good answers and the number of expected answers (i.e. present in reference)

$$R = \frac{tp}{tp + fn} \tag{2}$$

- *fn*: False negatives: number of missed items (*deletions*)

Example 1

REF: <pers> Bertrand Delanoë </pers> was elected mayor of <loc> Paris </loc>

HYP1 : <pers> Bertrand Delanoë </pers> was elected <pers> mayor </pers> of <loc> Paris </loc>

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

$\rightarrow$ HYP1 produces **noise**

Example 2

REF: <pers> Bertrand Delanoë </pers> was elected mayor of <loc> Paris </loc>

HYP2: <pers> Bertrand Delanoë </pers> was elected mayor of Paris

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

$\rightarrow$ HYP1 produces **silence**

- Precision takes into account insertions and substitutions (false positives)
- Recall takes into account deletions (false negatives)

How to combine the two?

F-measure

F-measure, defined as the **harmonic mean between Precicion and Recall**:

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

Where β is a **weight** to adjust the importance of P ou R
(if 1, equally important).

Examples

REF: <pers> Bertrand Delanoë </pers> was elected mayor of <loc> Paris </loc>

HYP1 : <pers> Bertrand Delanoë </pers> was elected <pers> mayor </pers> of<loc> Paris </loc>

HYP2: <pers> Bertrand Delanoë </pers> was elected mayor of Paris

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

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

Inconvenience of classical measures

- Merging P and R minimize the weight of insertions and deletions compared to substitution, whatever β is [MKSW99]
- With fine-grained typologies, we need metrics that differentiate between error types.

Error types

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

HYP1: the <pers.ind> president </pers.ind> of Ford
→ boundary mistake

HYP2: the <pers.coll> president of Ford </pers.coll>
→ sub-type mistake

HYP3: the <pers.coll> president </pers.coll> of Ford
→ sub-type and boundary mistake.

Measures based on error types

- SER: Slor Error Rate
- ETER: Entity Tree Error Rate.

7. Zoom on DH

Performances of NE processing

NE processing is good when:

- language: English
- domain: news
- typology: simple

NE processing in DH: challenges

- **Noisy input**
  inherent in the source or from post-processing (OCR) e.g.
  Constap. iipopjle, Buch irest, M" Lucile
- **Language evolution** spelling variants, old naming conventions
  H???rnevi /_ Arnevi, Kallmar /_ Kallmar*
- **Domain specificity**
  - e.g. Vehicles (**VEH**) are important for event detection within works of fiction
- **Poor resource coverage**
  - low-resourced languages (e.g. Coptic, Ancient Greek, Old German)
  - minor or unknown entities, esp. for ORG type
  - lack of appropriate trigger words e.g. burgomaster, tailor, munition specialist

NE processing in DH: requirements

- ability to **(re-)train** with relatively smaller amount of data
- tools/models should be **robust** to noisy input
- adaptability to domain-specific tasks

NE processing in DH: applications

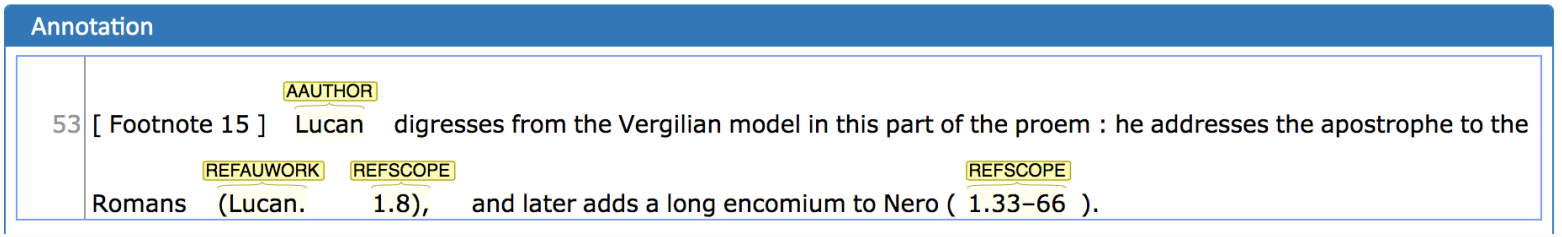
- **Enrichment** of museum/library catalogues
  - Europeana library metadata
  - US Holocaust Memorial Museum (EHRI project)
- Domain-specific adaptations
  - Annotations of fictional characters in literary texts
  - **Extraction of bibliographic references to classical texts**
  - Extraction of structured data from publications in the art conservation domain
  - ...
- Advanced text analysis platforms
  - ALCIDE (Analysis of Language and Content In a Digital Environment)
- Integration of NER software within **annotation platforms**
  - Recogito, via plugin system
  - INCEpTION, via APIs

Extraction of citations to primary sources in Classics

Example 1 (vgl. Hdt. 2, 170 f. u. 6. Apul. met. 11,23,9. Auson. Mos. 186 f. Ov. fast. 3, 325 f. 4,552. Verg. Aen. 6, 264ff. (E. Norden z. St.). Stat. Theb. 4, 516. Apoll. Rhod. 4, 249).

Example 2 “The picture of Achilles and of the Iliad that emerges from the twenty explicit references in the first half of the Aeneid is almost totally negative. Achilles is the unyielding (inmitis, 1 .30, 3.87), ferocious (saevus, 1.458, 2.29) warrior of Iliad 20 and 21 ; he is the preeminent killer of Trojans (1.30, 458, 468, 475, 484 ; 1.458-493, 2.196-198, 3.87, 5.803-811) and of Hector in particular (1.99, 483-484 ; 2.270-279 ; 6.168).”

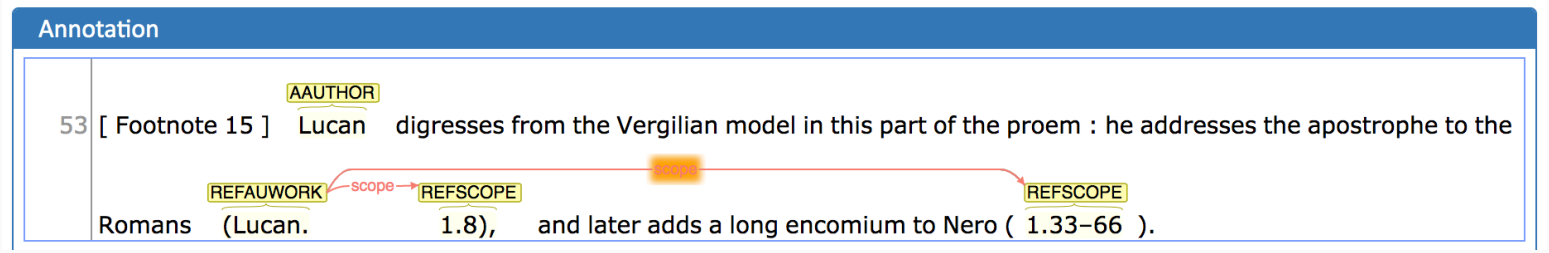
Recognition and Classification



Entities

- **aauthor** Homer
- **awork** Georgics
- **refauwork** Lucan.
- **scope** 11,4,11 ; IX 49

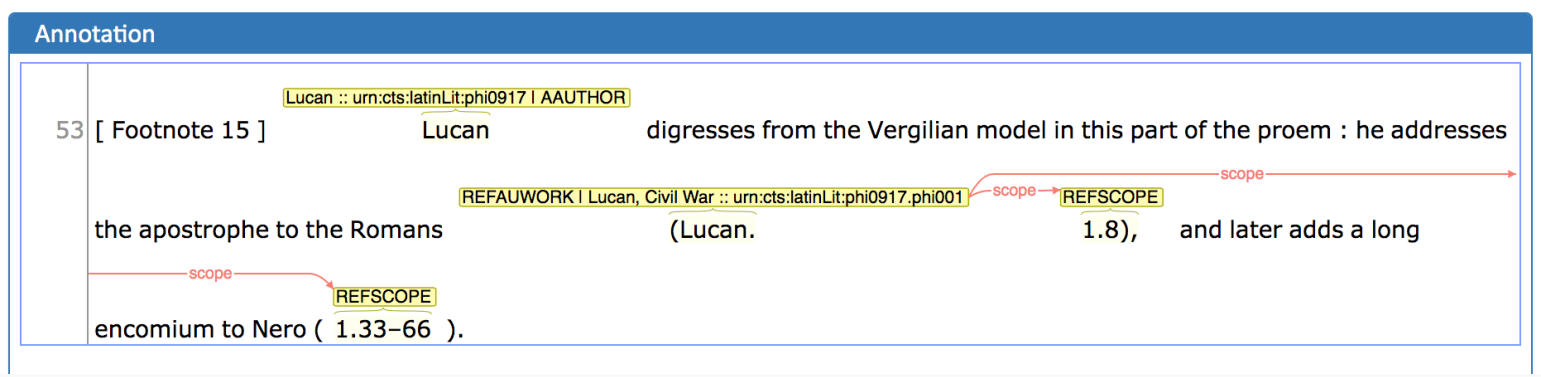
Relation Extraction



Possible combinations

- **refauwork + scope** Pliny, nat. 11, 4, 11
- **aauthor + scope** Ammianus (15, 8, 7)
- **awork + scope** Trabajos 159–173

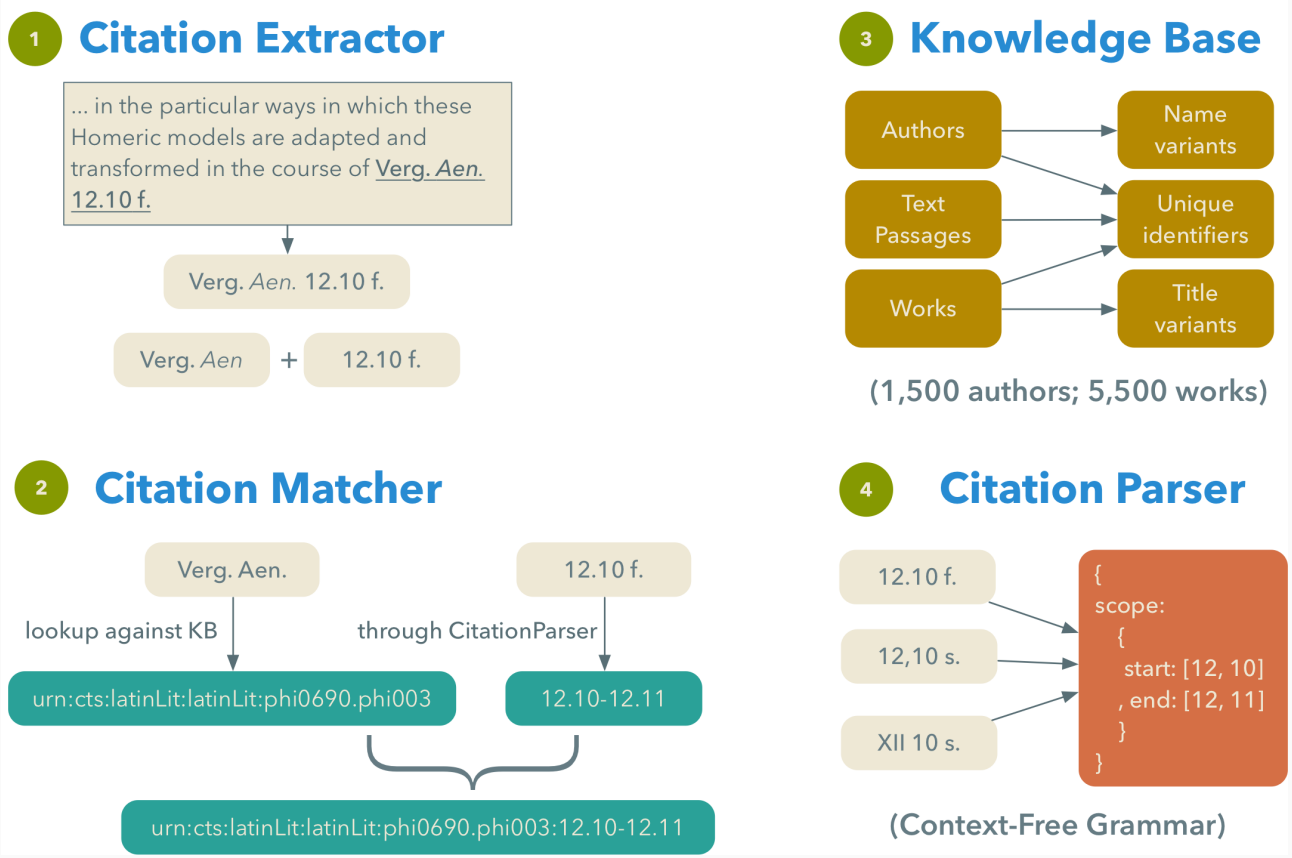
Entity/Relation Disambiguation



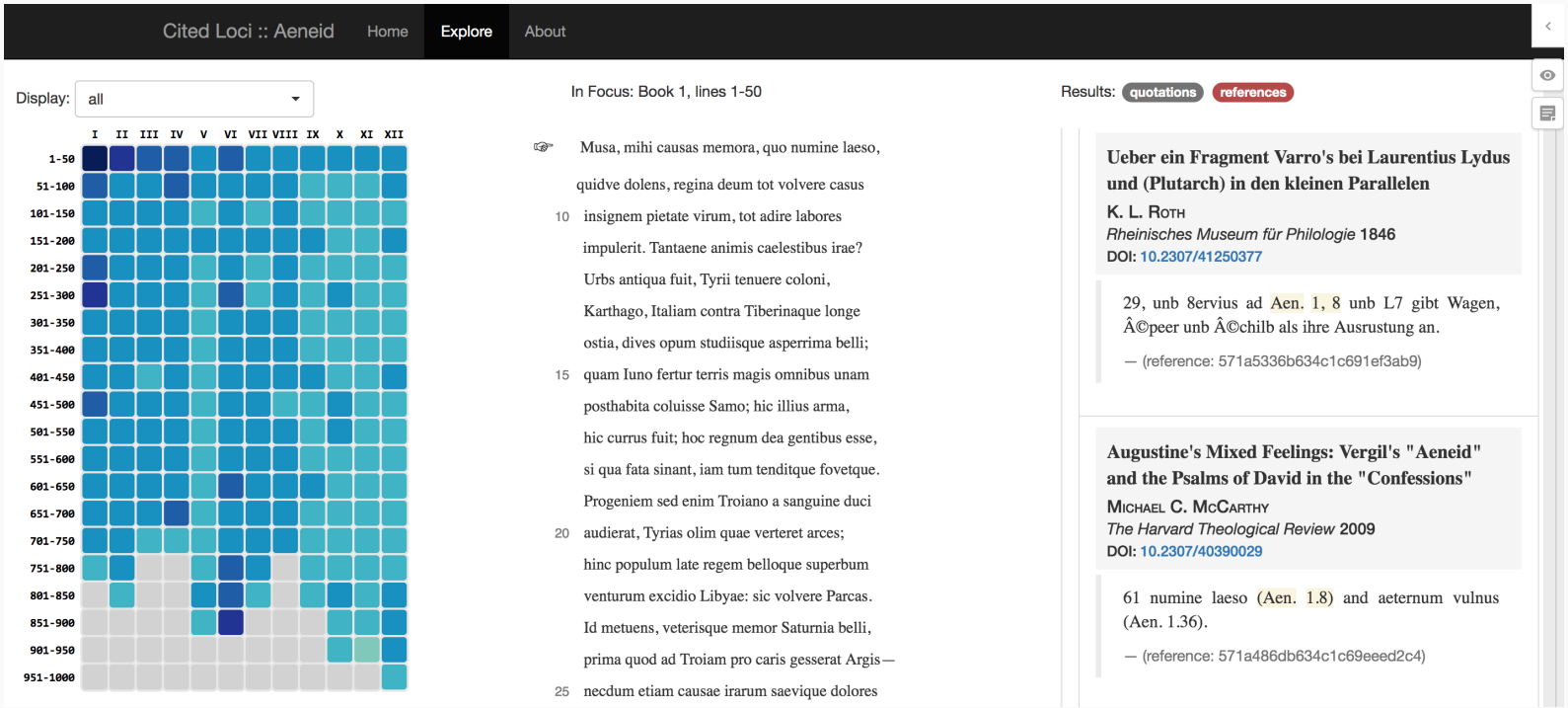
Linking via CTS URNs

- Lucan `urn:cts:latinLit:phi0917.001`
- Lucan. 1.8 `urn:cts:latinLit:phi0917.001:1.8`
- Lucan. 1.33-66 `urn:cts:latinLit:phi0917.001:1.33-1.66`

Extraction pipeline



Semantic Indexing of Publications



http://aeneid.citedloci.org

Domain adaptation

- What are entities of interest? What is the goal?
- What is the best way to annotate these entities?
- Which resources exist/need to be created?
  - lexical resources (e.g. gazetteers)
  - knowledge bases
  - annotated corpora

Contacts

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

Matteo Romanello
DHLAB
matteo.romanello@epfl.ch

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

Swiss National Science Foundation - Grant CR-SII5_173719

References i

- Alan Akbik, Duncan Blythe, and Roland Vollgraf, *Contextual string embeddings for sequence labeling*, Proceedings of the 27th International Conference on Computational Linguistics (Santa Fe, New Mexico, USA), Association for Computational Linguistics, August 2018, pp. 1638–1649.
- Michele Banko, Michael J Cafarella, Stephen Soderland, Matthew Broadhead, and Oren Etzioni, *Open information extraction from the web.*, IJCAI, vol. 7, 2007, pp. 2670–2676.
- Caroline Brun and Caroline Hagege, *Intertwining deep syntactic processing and named entity detection*, Advances in Natural Language Processing, Springer, 2004, pp. 195–206.

References ii

- Christopher Colah, *Understanding lstm networks*, electronic `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`, 2015.
- M. Ehrmann, *Les entités nommées, de la linguistique au TAL: statut théorique et méthodes de désambiguisation*, Ph.D. thesis, Université Denis Diderot, Paris, 2008, (sous la direction de M. Bernard Victorri).
- Nathalie Friburger, *Reconnaissance automatique des noms propres: application à la classification automatique de textes journalistiques*, Ph.D. thesis, Tours, 2002.

References iii


William A Gale, Kenneth W Church, and David Yarowsky, *One sense per discourse*, Proceedings of the workshop on Speech and Natural Language, Association for Computational Linguistics, 1992, pp. 233–237.


Cyril Grouin, Sophie Rosset, Pierre Zweigenbaum, Karën Fort, Olivier Galibert, and Ludovic Quintard, *Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview*, Proceedings of the Fifth Linguistic Annotation Workshop (LAW-V) (Portland, OR), Association for Computational Linguistics, June 2011, pp. 92–100.


Klaus Greff, Rupesh K. Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber, *Lstm: A search space odyssey*, IEEE Transactions on Neural Networks and Learning Systems **PP** (2016), no. 99, 1–11.

References iv

- Aurélien Géron, *Hands-on machine learning with scikit-learn and tensorflow*, O'Reilly Media, 2017.
- Zhiheng Huang, Wei Xu, and Kai Yu, *Bidirectional LSTM-CRF models for sequence tagging*, CoRR **abs/1508.01991** (2015).
- Heng Ji, Xiaoman Pan, Boliang Zhang, Joel Nothman, James Mayfield, Paul McNamee, Cash Costello, and Sydney Informatics Hub, *Overview of tac-kbp2017 13 languages entity discovery and linking.*, TAC, 2017.
- Andrej Karpathy, *The unreasonable effectiveness of recurrent neural networks*, electronic `http://karpathy.github.io/2015/05/21/rnn-effectiveness/`, 2015.

References v

- Omer Levy and Yoav Goldberg, *Linguistic regularities in sparse and explicit word representations*, Proceedings of the Eighteenth Conference on Computational Natural Language Learning (Ann Arbor, Michigan), Association for Computational Linguistics, June 2014, pp. 171–180.
- David McDonald, *Internal and external evidence in the identification and semantic categorization of proper names*, Corpus processing for lexical acquisition (1996), 21–39.
- John Makhoul, Francis Kubala, Richard Schwartz, and Ralph Weischedel, *Performance measures for information extraction*, In Proceedings of DARPA Broadcast News Workshop, 1999, pp. 249–252.
- Chris Potts, *Distributional approaches to word meanings*, 2013.

References vi

- Xin Rong, *word2vec parameter learning explained*, CoRR **abs/1411.2738** (2016).
- M Sahlgren, *The distributional hypothesis*, Italian Journal of Linguistics **20** (2008), no. 1, 33–54.
- C E Shannon, *Prediction and Entropy of Printed English*, Bell Systems Technical Journal **30** (1951), no. January, 50–64.
- Mihai Surdeanu and Heng Ji, *Overview of the english slot filling track at the tac2014 knowledge base population evaluation*, Proc. Text Analysis Conference (TAC2014), 2014.
- Richard Socher and Christopher D. Manning, *Deep learning for nlp (without magic)*, 2013.

References vii

- Richard Socher, *Cs224d deep learning for natural language processing lecture 2: Word vectors*, electronic`http://cs224d.stanford.edu/lectures/CS224d-Lecture2.pdf`, 2016.
- Diana Santos, Nuno Seco, Nuno Cardoso, and Rui Vilela, *HAREM: An Advanced NER Evaluation Contest for Portuguese*, lrec (Genoa), May 2006, pp. 1640–1643.
- Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata, *Extended named entity hierarchy*, Proceedings of the Third International Conference on Language Resources and Evaluation, LREC 2002, May 29-31, 2002, Las Palmas, Canary Islands, Spain, 2002.

References viii

- Beth M. Sundheim, *Overview of the third message understanding evaluation and conference*, THIRD MESSAGE UNDERSTANDING CONFERENCE (MUC-3): Proceedings of a Conference Held in San Diego, California, May 21-23, 1991, 1991.
- Erik F. Tjong Kim Sang and Fien De Meulder, *Introduction to the conll-2003 shared task: Language-independent named entity recognition*, Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4 (Stroudsburg, PA, USA), CONLL '03, Association for Computational Linguistics, 2003, pp. 142–147.