Natural Language processing (NLP)
Named entity recognition

Thierry Poibeau (CNRS & PSL/ENS, Lattice)

thierry.poibeau@ens.psl.eu

DHAI, 29 March 2022

Named Entity Recognition (NER)

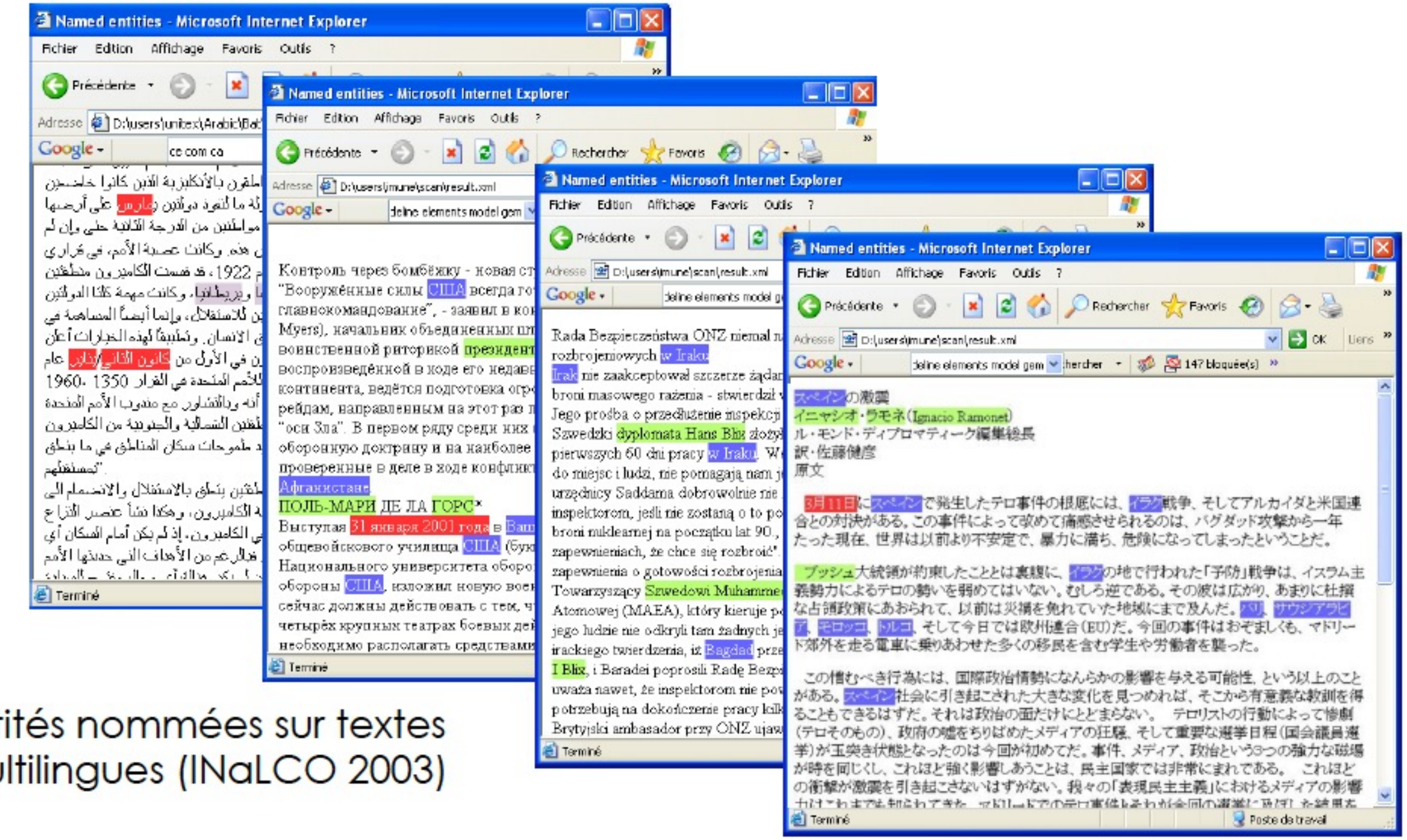
- A subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations
- By extension, NER may also include medical codes, time expressions, quantities, monetary values, percentages, etc.

Apollinaire’s peregrinations in « *Le passant de Prague* »

“*Voilà ! J’avais eu affaire, **rue de la Pépinière**, près de la **place Saint-Augustin**, et je revenais par le **boulevard Malesherbes** en l’intention de prendre l’omnibus à la **Madeleine**. Tout à coup, au coin de la **rue des Mathurins**, un homme se dressa devant moi en criant : “Madame ou mademoiselle, [...]. ”.*”

(Frontini et al 2016)
https://hal.archives-ouvertes.fr/hal-01363709

Multilingual NER



Entités nommées sur textes multilingues (INaLCO 2003)

Definition and Typology of Named Entities
Automatic Recognition
Conclusion

Definition and Typology of Named Entities
Automatic Recognition
Conclusion

Definition Problems

- NER definition and categories depends on the task / the application
  - Ex. In biology, genes and proteins can be seen as named entity
  - (Named) entity = entity of interest
- Common problems
  - Lexical ambiguity: *Paris, Texas* vs *Paris, France* vs *Paris Hilton*
  - One entity, different names: *Paris, Paname, la capitale*
  - Metonymy and categorial ambiguity: *Paris is claiming leadership as a nuclear power*

Definition Problems

- Named entities, definite descriptions, coreference
  - Macron … The president of France … He…
- Positions that depends on time
  - The president of France, in 2020, in 2005, in 1995?
- Groups of people, poorly delimited nouns
  - The North of Portugal, The (former) organizing Olympic committee
- Various things / objects
  - God, names referring to real people in novels, or imaginary people in the news (*Mickey*)

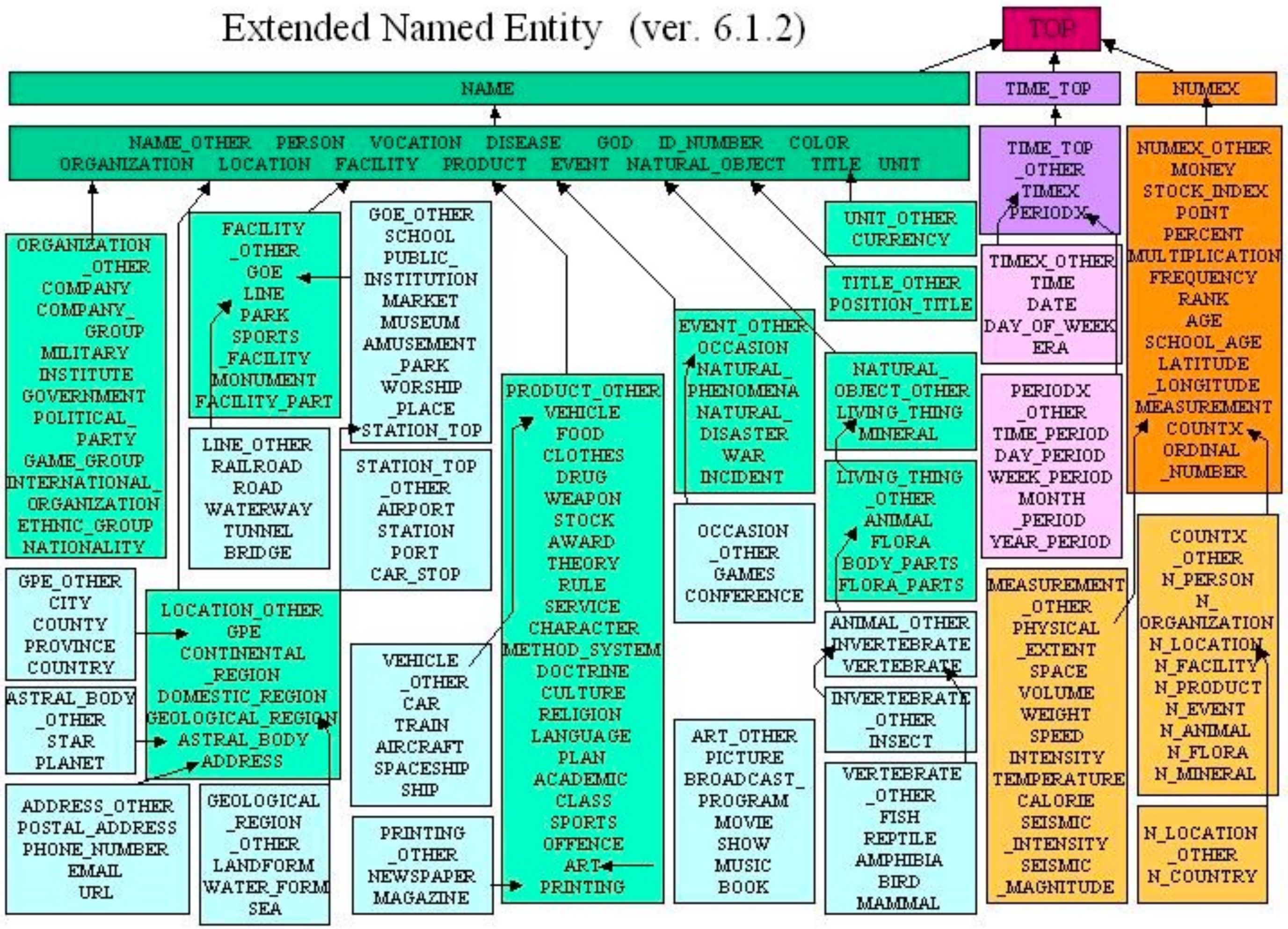
NER Typology

- First typology defined for the Message Understanding Conferences (MUC-6, MUC-7, 1995-1998)
- 7 main types and only 3 'real' named entity types (ENAMEX : PERS, ORG, LOC)

ENAMEX	TIMEX	NUMEX
PERSON	DATE	PERCENT
ORGANIZATION	TIME	MONEY
LOCATION		

Types	Exemple	Contre-exemple
ORG	*DARPA*	our university
PERS	*Harry Schearer*	St. Michael
LOC	*U.S.*	53140 Gatchell Road
MONEY	*19 dollars*	en dollars ? ça fait 19
TIME	*8 heures*	la nuit dernière (*)
DATE	le *23 juillet*	en **juillet** dernier (*)

The Extended NER Typology (Sekine, 2002)



Annotation Comparison

Phrase	MUC-7	CoNLL03	ACE	Disag.
Baltimore defeated the Yankees	<Baltimore>LOC <Yankees>ORG (ref. A.1.6)	<Baltimore>ORG <Yankees>ORG	<Baltimore>NAM.ORG.SPO <Yankees>NAM.ORG.SPO (ref. 6.2)	C
Zywiec Full Light	<Zywiec>ORG ("Full Light" no markup, ref. A.1.7)	<Zywiec>ORG <Full Light>MISC	<Zywiec>NAM.ORG (ref. 9.3.2)	I, C
Empire State Building	no markup (ref. 4.2.3)	<Empire State>LOC	<Empire State Building> NAM.FAC.Building (ref. 9.3.2)	I, C, B
Alpine Skiing-Women's World Cup Downhill	no markup (ref. A.2.4)	<World Cup>MISC (ref. guide)	<Women>NOM <World>NOM (ref. 9.3.3)	I, C, B
the new upper house of Czech parliament	<parliament>ORG (ref. A.4.3, A.1.5)	<Czech>LOC	<Czech parliament>NOM (ref. 9.3.2)	I, C, B
Stalinist nations	no markup (ref. A.1.6)	<Stalinist>MISC	no markup (ref. 5.2.1)	I
Wall Street Journal	no markup (ref. A.1.7)	<Wall Street Journal>ORG	<Wall Street Journal> NAM.ORG.MED (ref. 9.5.3)	I

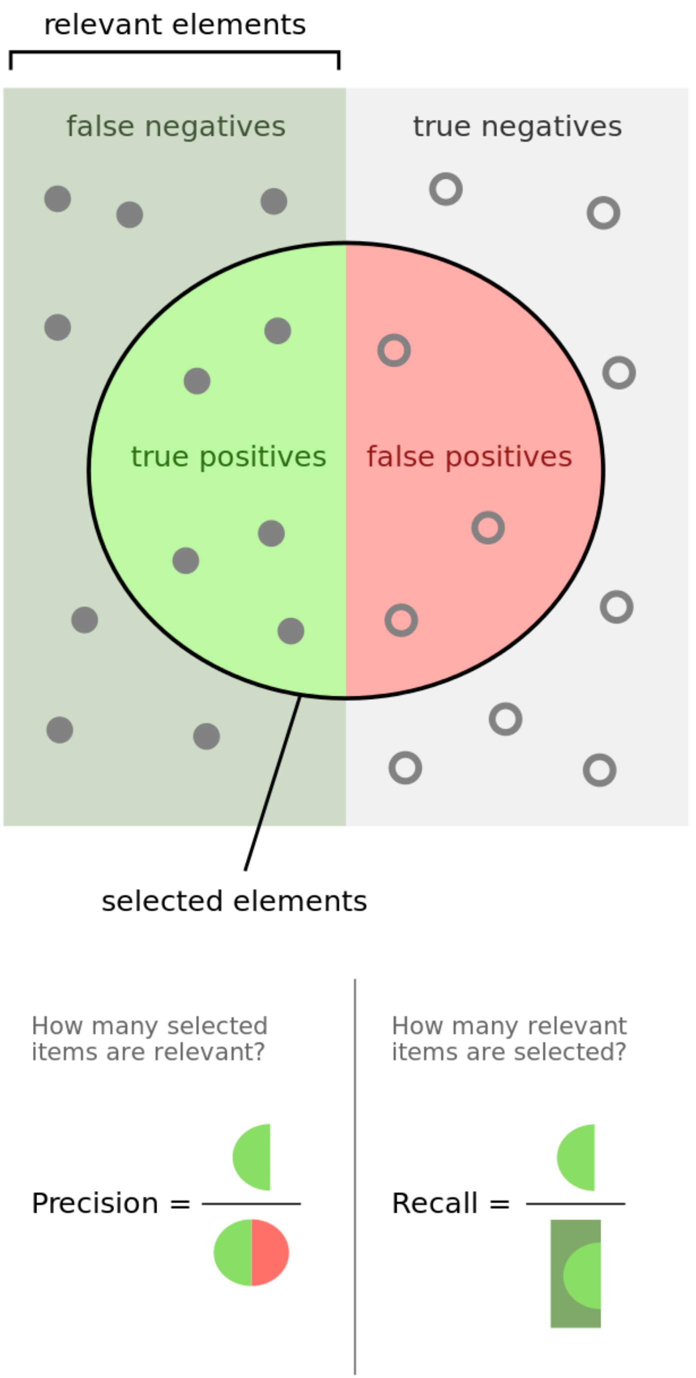
ESTER2, NER in French (2008-2009)

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

(a) Le [ent=org.pol-] *Parti Communiste* [-ent=org.pol] a peu de chance d'être au second tour.
(b) Le [ent=org.pol-] *RPR* [-ent=org.pol] est dissous en 2002.
(c) La course à la [ent=org.pol-] *Mairie de [ent=loc.admi-] Paris [-ent=loc.admi]* [-ent=org.pol] a commencé entre les deux principaux candidats.
(d) La [ent=org.pol-] *CIA* [-ent=org.pol] est chargée de l'acquisition du renseignement à l'étranger.
(e) Pendant la Guerre froide, le [ent=org.pol-] *KGB* [-ent=org.pol] joua un rôle crucial dans la survie de l'[ent=org.gsp-] État soviétique [-ent=org.gsp]

Evaluation: Precision and Recall

- For any element in the dataset
  - Have all the relevant elements been tagged?
  - Are all the tagged elements relevant?
- Evaluation indicators
  - Precision: # relevant tagged element /# tagged elements
  - Recall: # relevant tagged element /# elements to be tagged
  - F-measure = 2 * (P * R) / P + R



Source : Wikipedia, https://en.wikipedia.org/wiki/Precision_and_recall

Inter Annotator Agreement

- Inter-annotator agreement is a measure of how well two (or more) annotators can make the same annotation decision for a certain category

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

where *Pr(a)* is the relative observed agreement among raters, and *Pr(e)* is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly seeing each category. If the raters are in complete agreemnt then $\kappa = 1$.

Definition and Typology of Named Entities

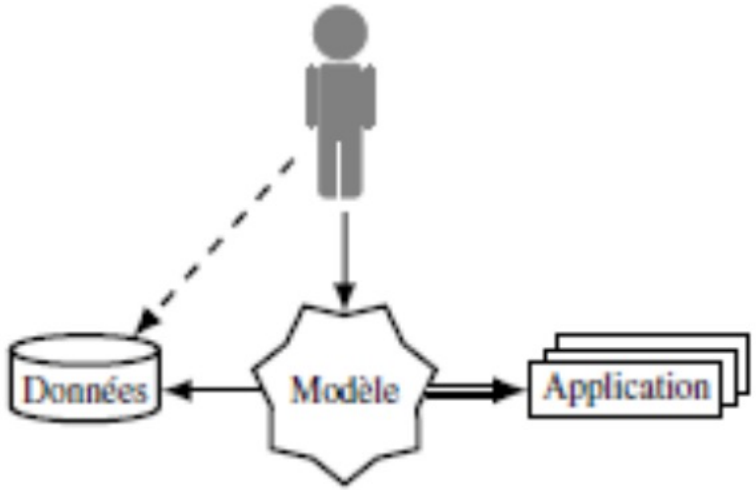
Automatic Recognition

Conclusion

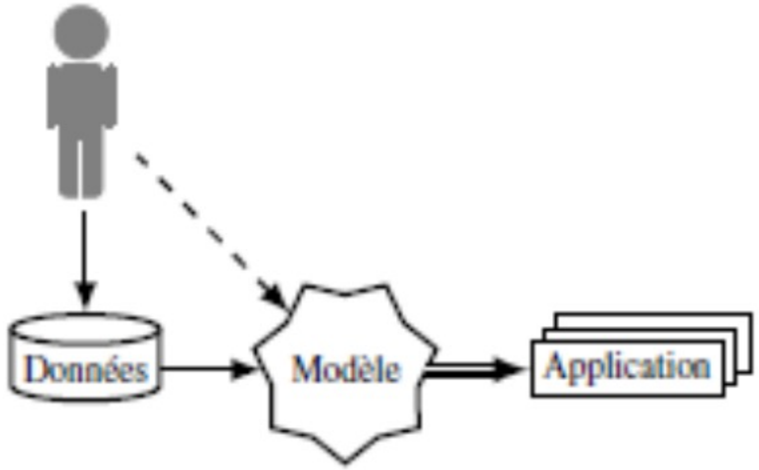
Main Features for NER

- Case (word begins with a capital letter, *Paris*)
- Ponctuation (word includes some punctuation, *S.N.C.F*)
- Number (word includes a number, *W3C*)
- Morphology (*Cambridgeshire, Oxfordshire*)
- Trigger (*Mr Bean, Ms. Jones, in Madrid*)
- Pos (gazetteers, lists of proper names)
- Unknown words (word not included in a general dictionary)
- Other kinds of contexts

Symbolic vs ML approaches



Système symbolique

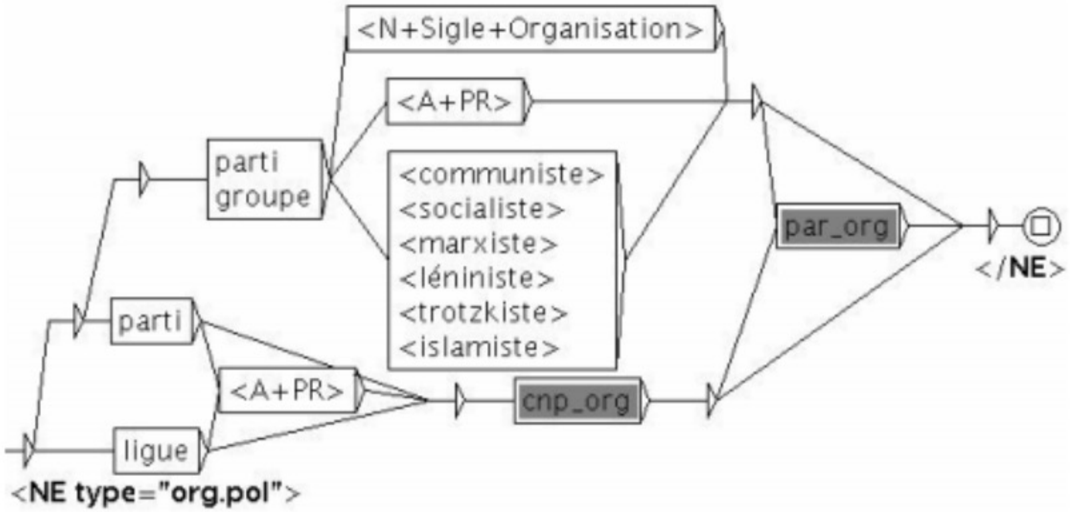


Système guidé par les données

Symbolic Approach

- Generally two main components
  - Dictionaries
  - Grammar

- Generally precise and accurate
- Hard to reach a good recall
- Maintenance problems



Nouvel et al., 2010

NE Annotation

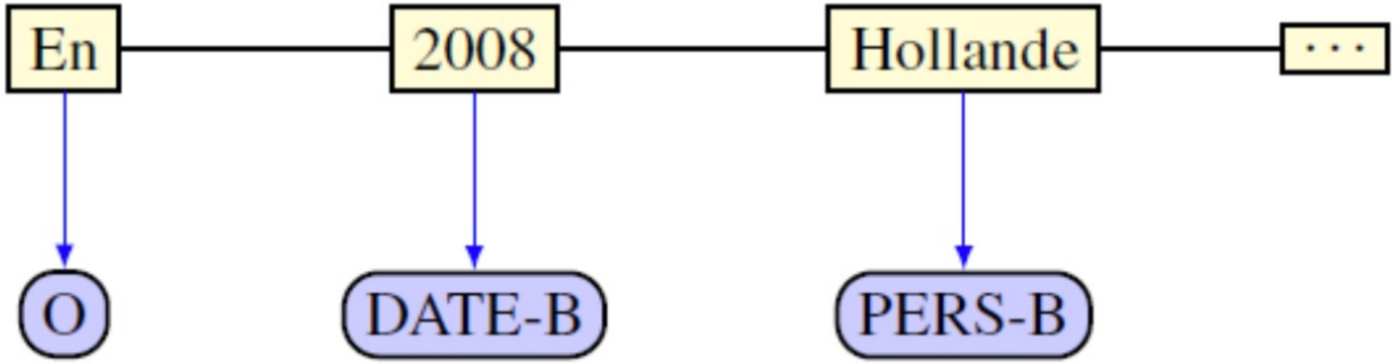
- BIO annotations

Mot	Catégorie
Lucy	PER
qui	O
descend	O
...	O
dit	O
la	PER
Faloise	PER
à	O
Fauchery	PER

Mot	Catégorie
Lucy	B-PER
qui	O
descend	O
...	O
dit	O
la	B-PER
Faloise	I-PER
à	O
Fauchery	B-PER

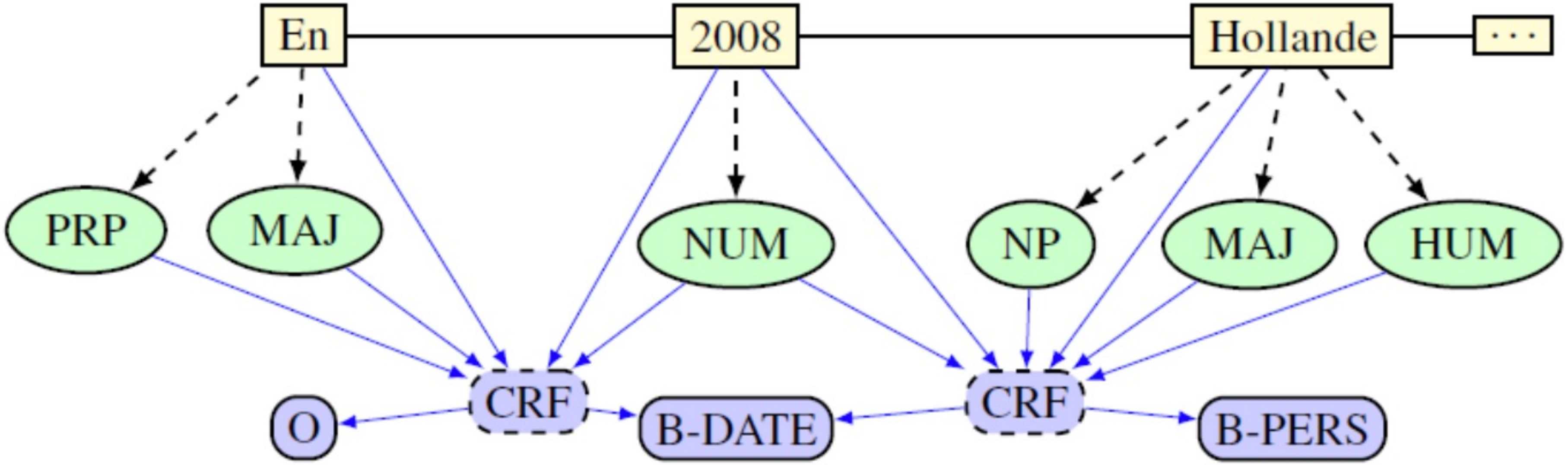
Baseline, with no context

- For each word, select the most probable tag
- Apply this to the text (in a purely procedural manner)



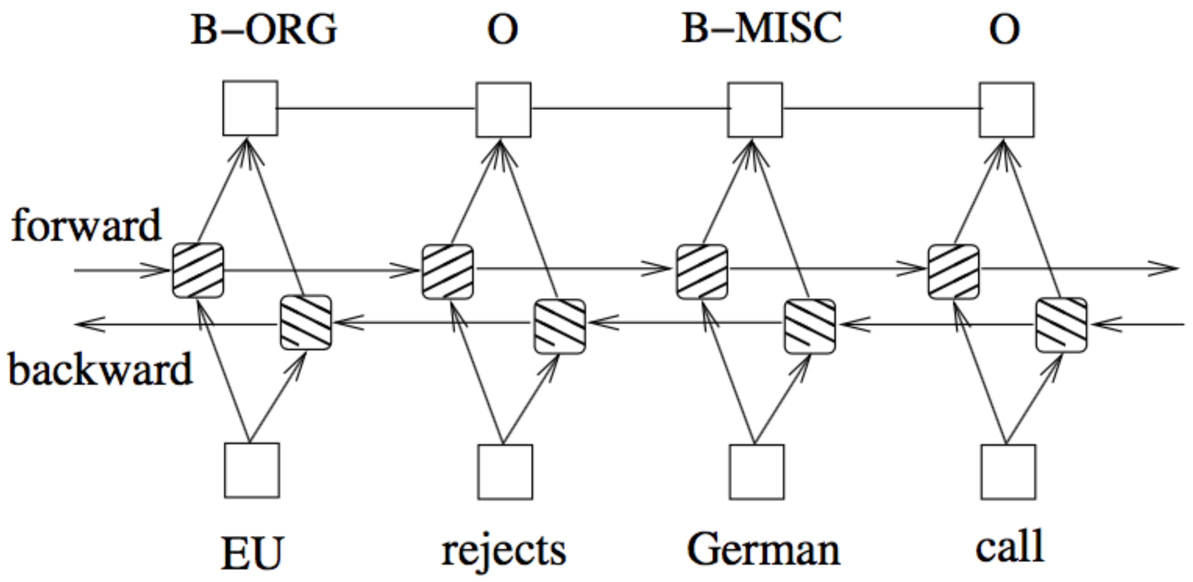
CRF, to take into account the context

- Conditional random Fields takes into account the context, as well as potential dependencies between tagged elements



LSTM

- Long short term memory networks
- Double chaining, from left to right, and from right to left



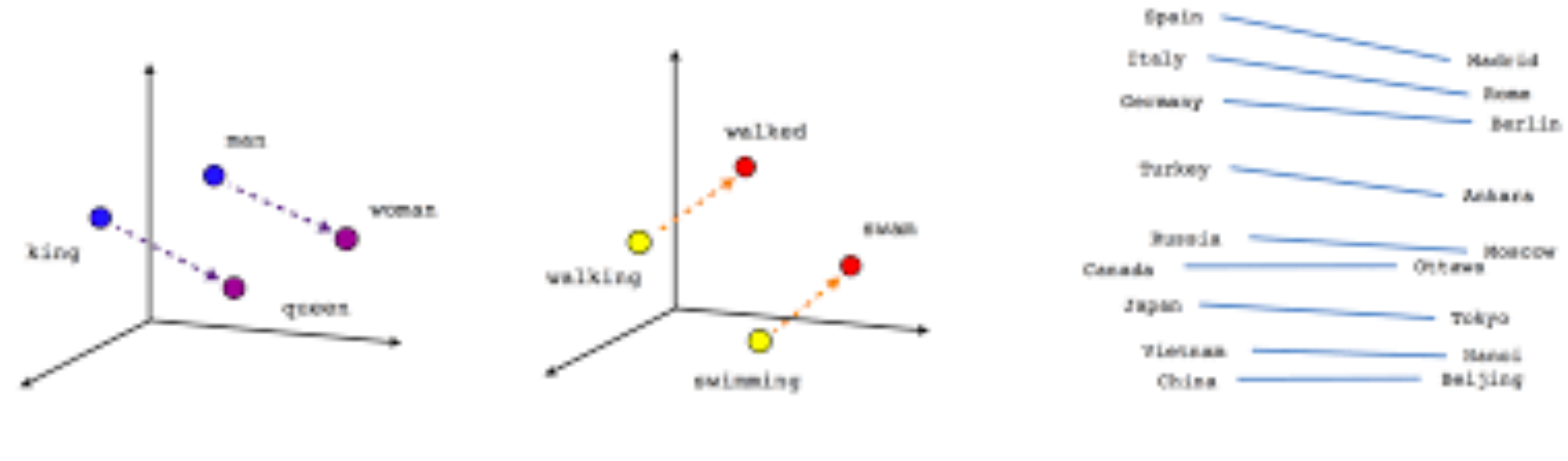
Deep learning and Language Models

- Most of the linguistic information is stored in a language models like Bert
- At the core of these models stands the notion of word embedding: a dense representation of the meaning of words by use of vectors
- Lexical proximity can be inferred from the vector representation

	Man (5391)	Woman (9853)	King (4914)	Queen (7157)	Apple (456)	Orange (6257)
Gender	-1	1	-0.95	0.97	0.00	0.01
Royal	0.01	0.02	0.93	0.95	-0.01	0.00
Age	0.03	0.02	0.7	0.69	0.03	-0.02
Food	0.04	0.01	0.02	0.01	0.95	0.97
⋮						

Language Models

- Language Models provide a very accurate representation of meaning and context
- Can be applied to a wide variety of tasks (cf. Glue, Superglue)
- Including NER



Male-Female   Verb tense   Country-Capital

Definition and Typology of Named Entities
Automatic Recognition
Conclusion

Entity Linking

- Establish a link between mentions of ENs in a text and the outer world.
- Disambiguation
  - "Goncourt" - Edmond de Goncourt or Jules de Goncourt ?

  « Voltaire", "François-Marie Arouet" Two different denomination dor the same person
- Link with an ontology (or a formal representation of a domain)
  - "Voltaire" – associated with his Wikipedia entry
    Thus, there are two goals:

Challenges (beyond NER)

- Cover more languages
- Reduce the amount of data required for learning
- Have more robust systems (from one domain to the other)