COST is supported by the EU Framework Programme Horizon 2020

Named Entity Recognition systems

carmen.brando@ehess.fr

Distant Reading

COST Action “Distant Reading” (COST Action CA16204) training school
March 22, 2021

Plan

- **Tool pipelines for linguistic analysis and NER systems**
- Challenges for NER
- Features for NER systems
- Types of NER systems
- Manual annotation & evaluation and training of NER systems
- Some available out-of-the-box NER systems
- Output NE annotation formats

Named entity recognition (NER)

…
I was born at Blunderstone, in
Suffolk, or 'there by', as they say
in Scotland.
…
I remarked that,
once or twice when Mr. Quinion
was talking, he looked at Mr.
Murdstone sideways
...

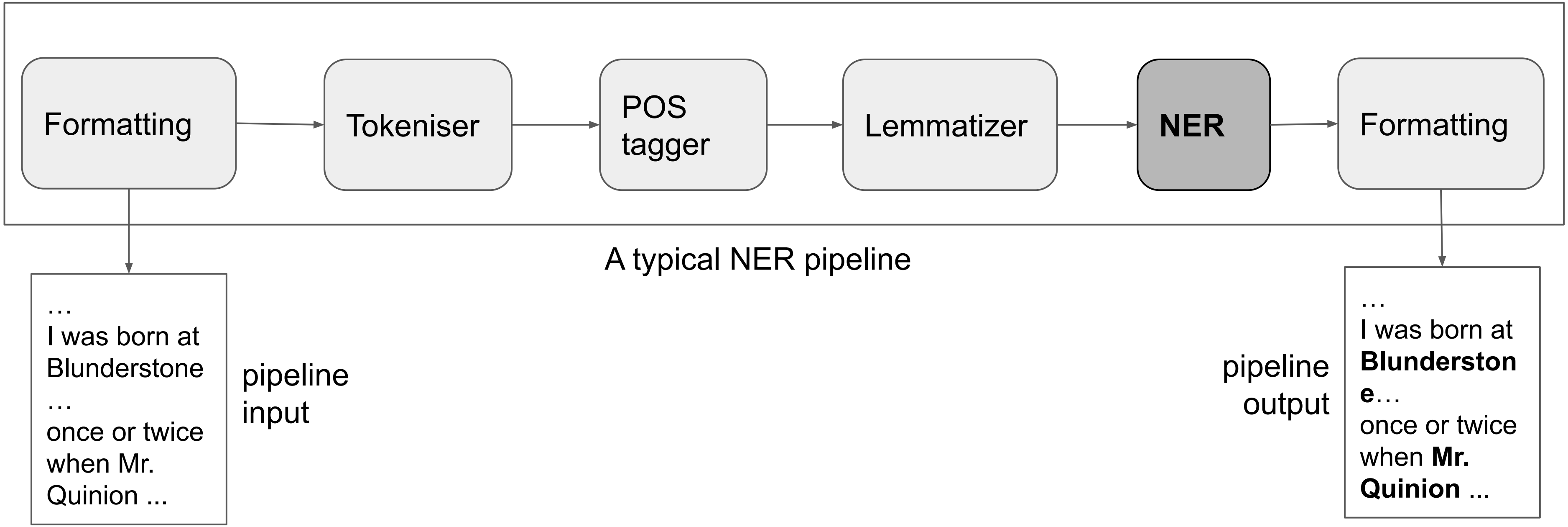
Raw text
(excerpt)

…
I was born at **<u>Blunderstone</u>**, in
<u>Suffolk</u>, or 'there by', as they
say
in **<u>Scotland</u>**.
…
I remarked that,
once or twice when **<u>Mr. Quinion</u>**
was talking, he looked at **<u>Mr.</u>**
<u>Murdstone</u> sideways
...

Annotated text in NE

- **Spotting**: determining the boundaries of the EN, i.e. which text segments are concerned
- **Classifying**: determining the type of the EN from a predefined typology (usually, LOC, PER, ORG)

Tool pipelines for linguistic analysis and NER systems



A typical NER pipeline

Several configurations are possible and depending on the language, some pipeline components may be **optional**, except from tokenisation. Sentence segmentation maybe be other component.

Plan

- Tool pipelines for linguistic analysis and NER systems
- **Challenges for NER**
- Features for NER systems
- Types of NER systems
- Manual annotation & evaluation and training of NER systems
- Some available out-of-the-box NER systems
- Output NE annotation formats

Challenges for NER: types of ambiguities

- Same name for several entities: *Paris* (France) and *Paris* (Texas)
- An entity may have several names: *Paris*, *Paname*
- Several categories possible for a NE (methonomy)
  - la *France*, … depending of the context, it may be an organisation or a location
  - "Le *prix Nobel de la Paix* s'est montré digne devant une telle épreuve" (language dependent)

Challenges for NER: difficulties in defining NE

- NE boundaries, different levels of granularity, for instance, la rue de Strasbourg, the executive committee of the Union of European Football Association
- Nested EN annotation, the Queen of England, la chapelle de la Vierge Marie
- fuzzy, collective or historical NE referent: the coasts of Guyana, Northern Europe, La Bohême

Plan

- Tool pipelines for linguistic analysis and NER systems
- Challenges for NER
- **Features for NER systems**
- Types of NER systems
- Manual annotation & evaluation and training of NER systems
- Some available out-of-the-box NER systems
- Output NE annotation formats

Features for NER systems

- Word-level features
- List lookup features
- Document and corpus features

Nadeau et Sekine 2007

Features for NER systems: word-level

- **Case**: - starts with a capital letter
- **Punctuation**: - Internal apostrophe, hyphen or ampersand (par ex, O'Connor)
- **Character**: - Possessive mark (ex : Esther's family)
- **Morphologie**: - Prefix, suffix, singular version, stem - Common ending
- **Part-of-speech**: - proper name, verb, noun, foreign word - recurring combination of categories, ex : madame François -> commun name + proper name

Nadeau et Sekine 2007

Features for NER systems: word-level (2)

- **NE Context**
  - local: words that precede or follow the EN, e.g. "I dislike Holland in Spiderman" vs. "His trip to Holland went well"
  - sometimes need wider context (sentence, close sentence), e.g. "I read up on Washington for my work"
- Contextual cues complement the indicators presented above

Features for NER systems: list lookup

List → “gazetteer”, “lexicon” and “dictionary”

- **General list** : - General dictionary - Stop words (function words) - Capitalized nouns (e.g., January, Monday) - Common abbreviations
- **List of entities** : - Organization, government, airline, educational - First name, last name - Astral body, continent, country, state, city
- **List of entity cues** : - Typical words in organization - Person title, name prefix, post-nominal letters - Location generic terms (road, bridge, ..), cardinal point

Nadeau et Sekine 2007

Features for NER systems: document/corpus level

- **Multiple occurrences** : - Other entities in the context - Uppercase and lowercase occurrences - Anaphora, coreference
- **Local syntax** : - Enumeration, apposition - Position in sentence, in paragraph, and in document
- **Corpus frequency** : - Word and phrase frequency - Co-occurrences - Multiword unit permanency

Nadeau et Sekine 2007

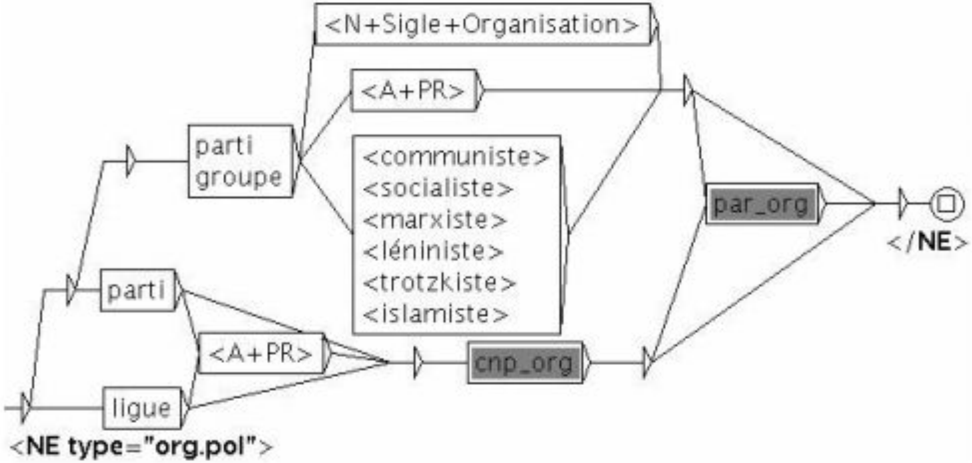
Plan

- Tool pipelines for linguistic analysis and NER systems
- Challenges for NER
- Features for NER systems
- **Types of NER systems**
- Manual annotation & evaluation and training of NER systems
- Some available out-of-the-box NER systems
- Output annotation formats

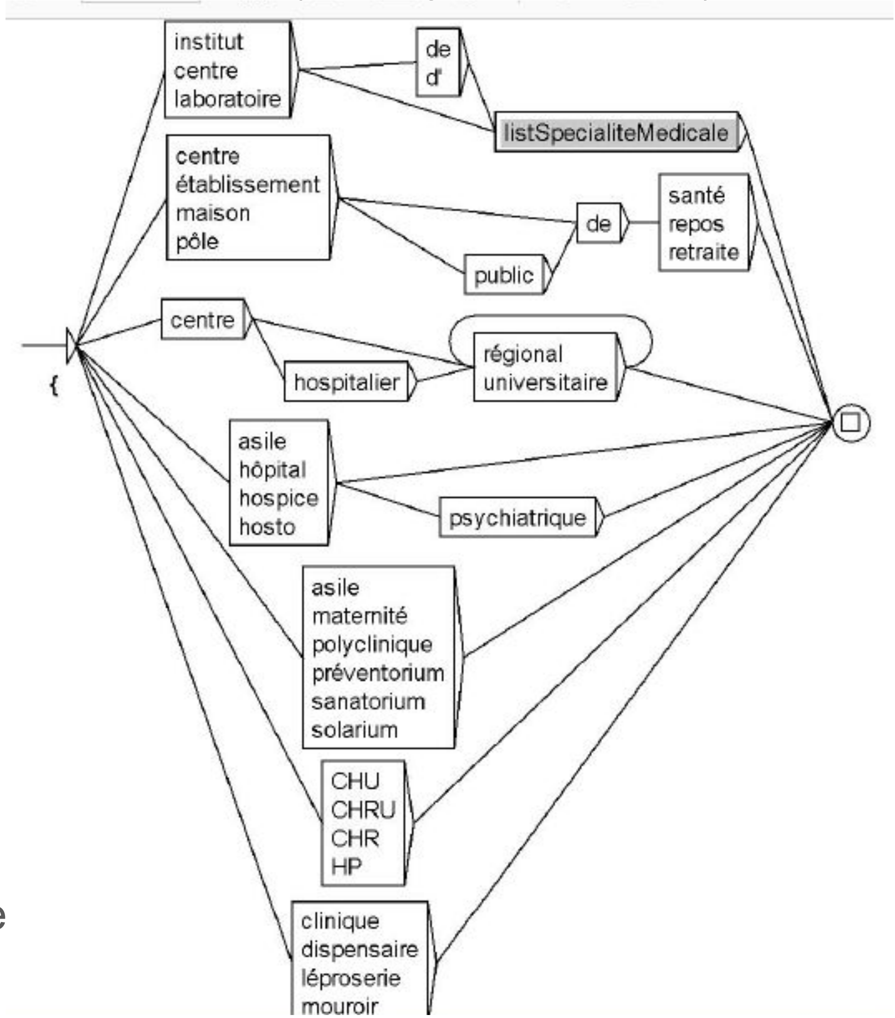
Types of NER systems

- **Approaches based on symbolic methods, based on rules developed by an expert and dictionaries (lists)**
- Statistical and data-driven approaches

Recognition of medical institutions (Fribourg et Maurel 2004)



Recognition of political organisations
(Nouvel et al 2010)

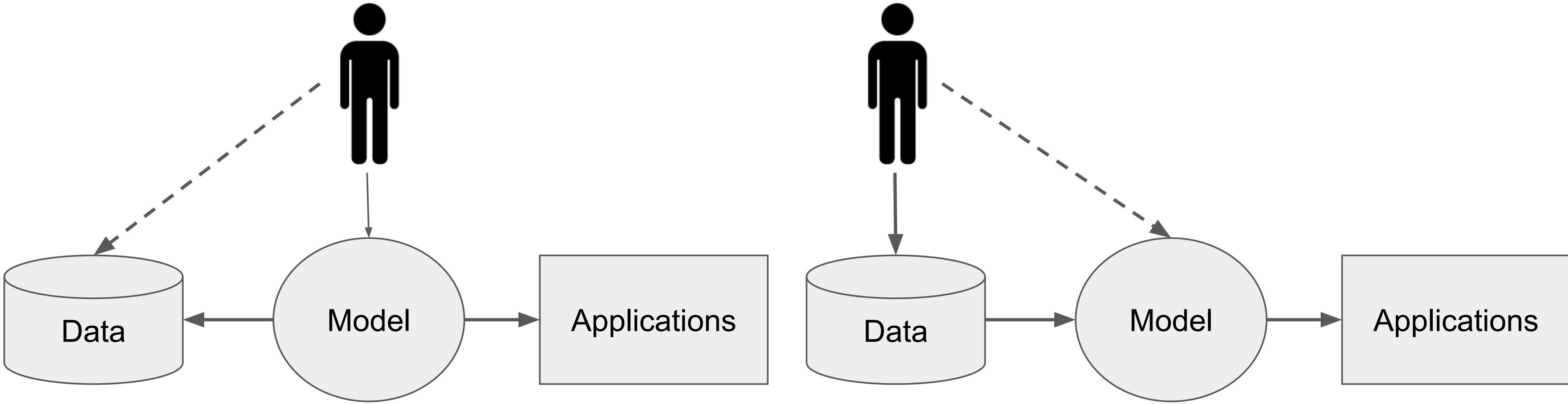


Transducers (special case of finite automata) in the form of Unitex graphs

Types of NER systems

- Approaches based on symbolic methods, based on rules developed by an expert and dictionaries (lists)
- **Statistical and data-driven approaches**

Statistical and data-driven approaches / rule-based systems



Symbolic systems

Statistical and data-driven approaches

interacts

visualise, evaluate, configure

(Nouvel et al 2015)

Statistical and data-driven approaches

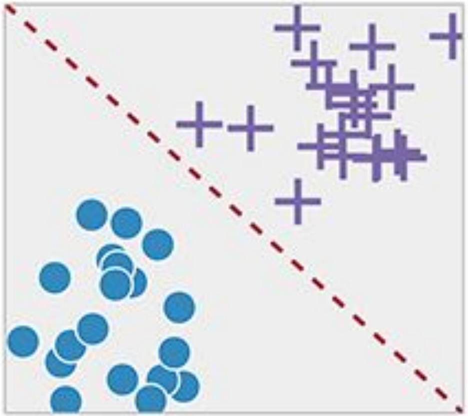
- **Supervised learning** rely on labelled data for training models,
- **Unsupervised learning** do not need labelled data
- **Semi-supervised learning** rely on both labelled and unlabelled data, and needs a small labelled data set to start the learning process
- **Neural network** based approaches, discover hidden features in the data by successive analysis of the text in layers, they may rely on unlabelled data for creating models which are afterwards fine-tuned for NER

Statistical and data-driven approaches

The way in which data is presented to the system is crucial because:

- the **quantity and quality** of the data can make the system more or less **accurate** (few false matches), **comprehensive** (few missed matches), **robust** (resistance to noise),
- The **type of text** on which the training is carried out conditions the **applicability of the model** to other types of text,
- **Pre-processing** (tokenisation, ..) and the way that **named entities are defined** can influence the way these are recognised.

NER is modelled as a classification problem



From data, the algorithm aims to determine **discrete values (categories)** to assign to a given **input word sequence** by calculating the decision by linear combination of samples.

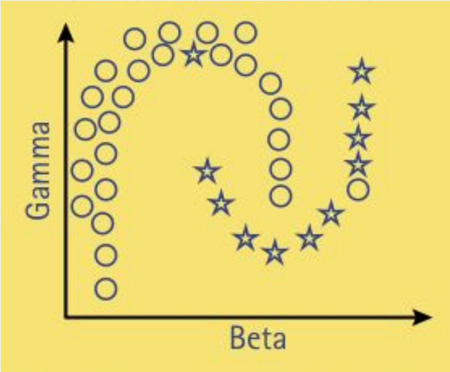
Token	**Category**
Lucy	B-PER
qui	O
descend	O
...	O
dit	O
la	B-PER
Faloise	I-PER
à	O
Fauchery	B-PER

Supervised learning

- These are classification systems that process a large annotated corpus and learn from examples of **texts annotated by humans**, so a model is trained,
- From the training corpora, these systems learn lists of entities and create disambiguation rules based on **discriminatory features**,
- **Conditional Markov fields (CRF)** are the most representative example of this type of approach, they take into account **the context of the word** to make decisions (a decision on a word in a text can influence the following decision)
- The **performance** of the NER :
  - depends on the **vocabulary transfer**, which is the proportion of words, without repetitions, appearing in the training and test corpus
  - is influenced by the **quantity** and **quality** of the annotated data and the **number of categories** to be learned (enough instances per class)

Neural networks based approaches

- In contrast to CRF models where the decision to label each word depends only on the words around it, these approaches take into account **all previously classified words**,
- Neural networks are an answer to the **limitations of linear classification**, have existed for several decades (the perceptron) but computer resources were limited at the time.



Neural networks based approaches (2)

- Examples of neural network architectures for NER: Long Short Term Memory (**LSTM**), **Bi-LSTM** (such as **ELMO**, **Flair**), **Bi-LSTM CRF**,
- The state of the art of NER is **BERT** (Bidirectional Encoder Representations from Transformers), they rely on **large neural networks trained** (on unlabelled data) on general tasks like language modeling and then **fine-tuned for classification task (NER)**

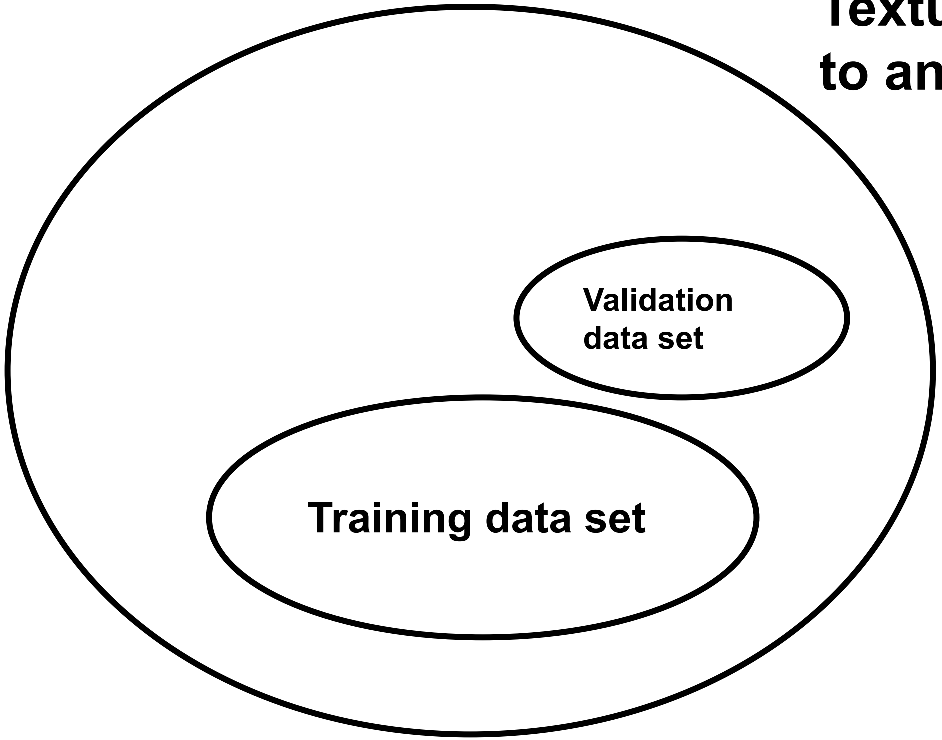
Plan

- Tool pipelines for linguistic analysis and NER systems
- Challenges for NER
- Features for NER systems
- Types of NER systems
- **Manual annotation, evaluation and training of NER systems**
- Some available out-of-the-box NER systems
- Output annotation formats

Training a NER system from user-annotated corpus: domain adaptation

Designing **NE annotation guidelines** is important to define **the application** and for the annotation of both **training and validation** data sets.

Textual corpora to analyse



Inter-annotator agreement to ensure consistency in the definition of NE

- It is a set of metrics to determine the consistency of annotations, as there is no "ground truth", the linguistic categories are therefore determined by **human judgement**,
- Once the corpus has been annotated, the quality and consistency of the annotations produced should be measured, i.e. to ensure that each annotator has had the **same understanding of the task and interpretation of the annotation guide**,
- **Cohen's Kappa** indicators for measuring expected agreement taking into account chance are the most widely used for two annotators. **Fleiss' Kappa** variant measures agreement in the presence of three annotators.

Evaluation of NER systems

gold: ***Phébus*** *parut en* ***Postillon de Lonjumeau*** *et* ***Minerve*** *en Nourrice normande.*

test: ***Phébus*** *parut en Postillon de* ***Lonjumeau*** *et* ***Minerve*** *en* ***Nourrice*** *normande.*

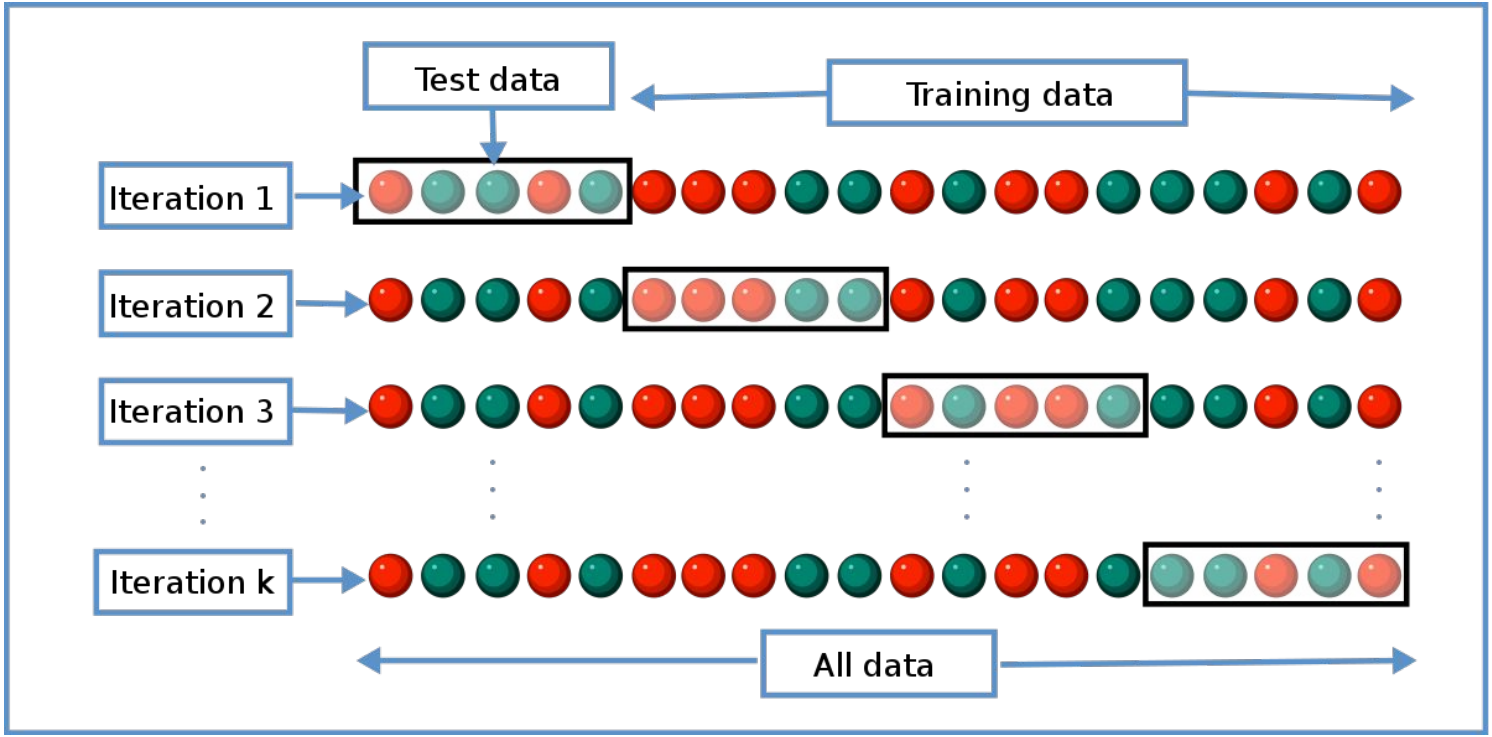
True positive (TP) False Negative (FN) TP False positive (FP)

- **Recall** : the number of correctly annotated entities wrt the total of manually annotated entities in the gold
  = VP / (VP + FN)

- **Precision** : the number of correctly annotated entities wrt the total of returned entities
  = VP / (VP + FP)

It is also possible to **extend** these measures to use **relaxed match** when comparing annotations, instead of **strict match** as presented above.

Choice of sub-corpora for training : Cross-validation approach



Plan

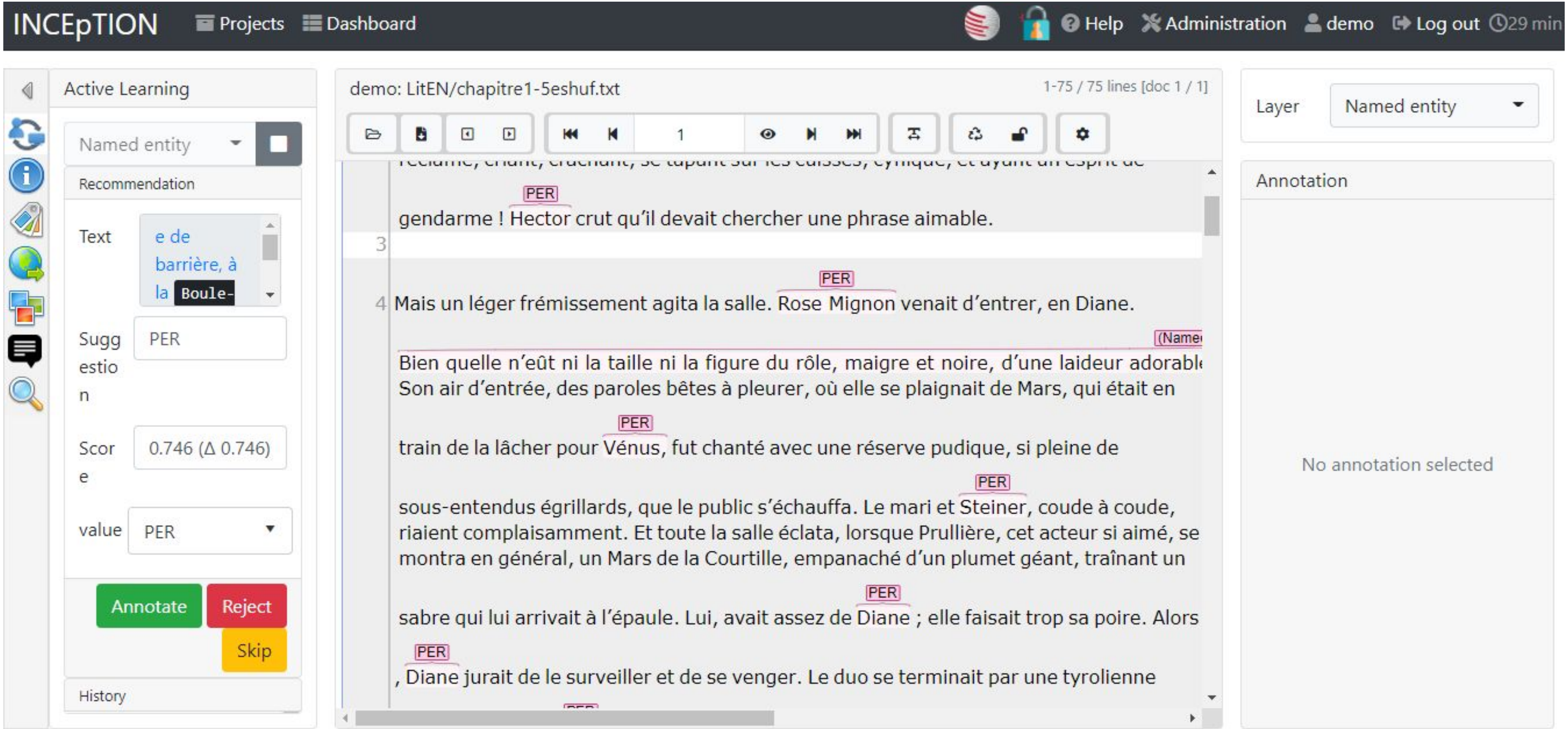
- Tool pipelines for linguistic analysis and NER systems
- Challenges for NER
- Features for NER systems
- Types of NER systems
- Manual annotation & evaluation and training of NER systems
- **Some available out-of-the-box NER systems**
- output NE annotation formats

Some available out-of-the-box NER systems

Many systems exist and most of them are specific to a language, other try to generalise to several languages. The most well-known systems are:

- Stanford NER, Stanza
- Spacy
- PALAVRAS-NER (for Portuguese)
- SEM (for French)
- Book NLP (a pipeline for literary text annotation, so far available for English)
- ...

Further information on Ranka’s presentation in the second part.



Inception (web interface assisted automatic annotation) : https://inception-project.github.io/

Plan

- Tool pipelines for linguistic analysis and NER systems
- Challenges for NER
- Features for NER systems
- Types of NER systems
- Manual annotation & evaluation and training of NER systems
- Some available out-of-the-box NER systems
- **output NE annotation formats**

Output NE annotation formats (1 token per line)

The CoNLL-2003 shared task data files contain four columns separated by a single space. Each word has been put on a separate line and there is an empty line after each sentence. Each line contains the following information separated by TAB: word, part-of-speech (POS) tag, syntactic chunk tag, **named entity tag**.

U.N.	NNP	I-NP	I-ORG
official	NN	I-NP	O
Ekeus	NNP	I-NP	I-PER
heads	VBZ	I-VP	O
for	IN	I-PP	O
Baghdad	NNP	I-NP	I-LOC

Extended versions of this format exist for including further kinds of linguistic annotations.

Output NE annotation formats (1 token per line)

BIO syntax does not permit any nesting, it is also necessary to distinguish a polylexical entity from two contiguous ENs of the same type.

Word	NE cat
expédition	O
vers	O
Tanger	B-LOC
...	O
le	B-PER
comte	I-PER
de	I-PER
Lambert-Sarrazin	I-PER
,	O

Format BIO

Word	NE cat
expédition	O
vers	O
Tanger	U-LOC
...	O
le	B-PER
comte	I-PER
de	I-PER
Lambert-Sarrazin	L-PER
,	O

Format BILOU (idem BIOES)

Output NE annotation formats (examples)

- XML based (inline,..)

```
<PERSON>Eliza</PERSON> and <PERSON>Georgiana</PERSON> had run for
<PERSON>Mrs. Reed</PERSON> , who was gone upstairs: she now came upon the
scene, followed by <PERSON>Bessie</PERSON> and her maid
<PERSON>Abbot</PERSON> .
```

- Standoff

```
T129    PER 15315 15320 Eliza
T131    PER 15846 15855 Georgiana
T132    PER 15873 15877 Reed
T133    PER 15943 15949 Bessie
T134    PER 15963 15968 Abbot
```

Some final remarks

- **Difficulty in adapting NER systems** to take into account texts that differ from the corpus for which the tool was designed, but increasingly, available corpora are multi-domain (Wikiner) and active learning can help to annotate enough data, but how many data is enough?
- EN categories available in existing annotated data are limited to the **three main categories** (PER, LOC, ORG) which are somehow different of those interesting for literary analysis, also their reuse to recognise more specific categories is limited,
- **few annotation guides focus on literary texts**, few systems suitable to deal with historical and poorly endowed languages and multilingual corpora,
- Further annotation layers are needed for a global analysis of the text: anaphora, coreference chains, object relations, sentiment analysis.