Natural Language Processing and Digital Humanities

Claire Gardent

CNRS/LORIA, Nancy, France

Empire Workshop, Reno
7-8 January 2020

Loria
Laboratoire lorrain de recherche
en informatique et ses applications

cnrs

UNIVERSITÉ
DE LORRAINE

Three ways of processing/visualising a text collection

Document-Based Visualisation

- Corpus ⇒ Graph, nodes = documents, edge = similarity
- Goal: Explore structure of document collection, Identify key topics, Weed out irrelevant documents

Location-Based Visualisation

- Corpus ⇒ Map, dots = locations + some information (text?)
- Goal: Geograhical representation of events

Event-Based Visualisation

- Corpus ⇒ Graph, nodes = entities, edge = relation
- Goal: Visualisation of the relations between entities (Social Netwok)

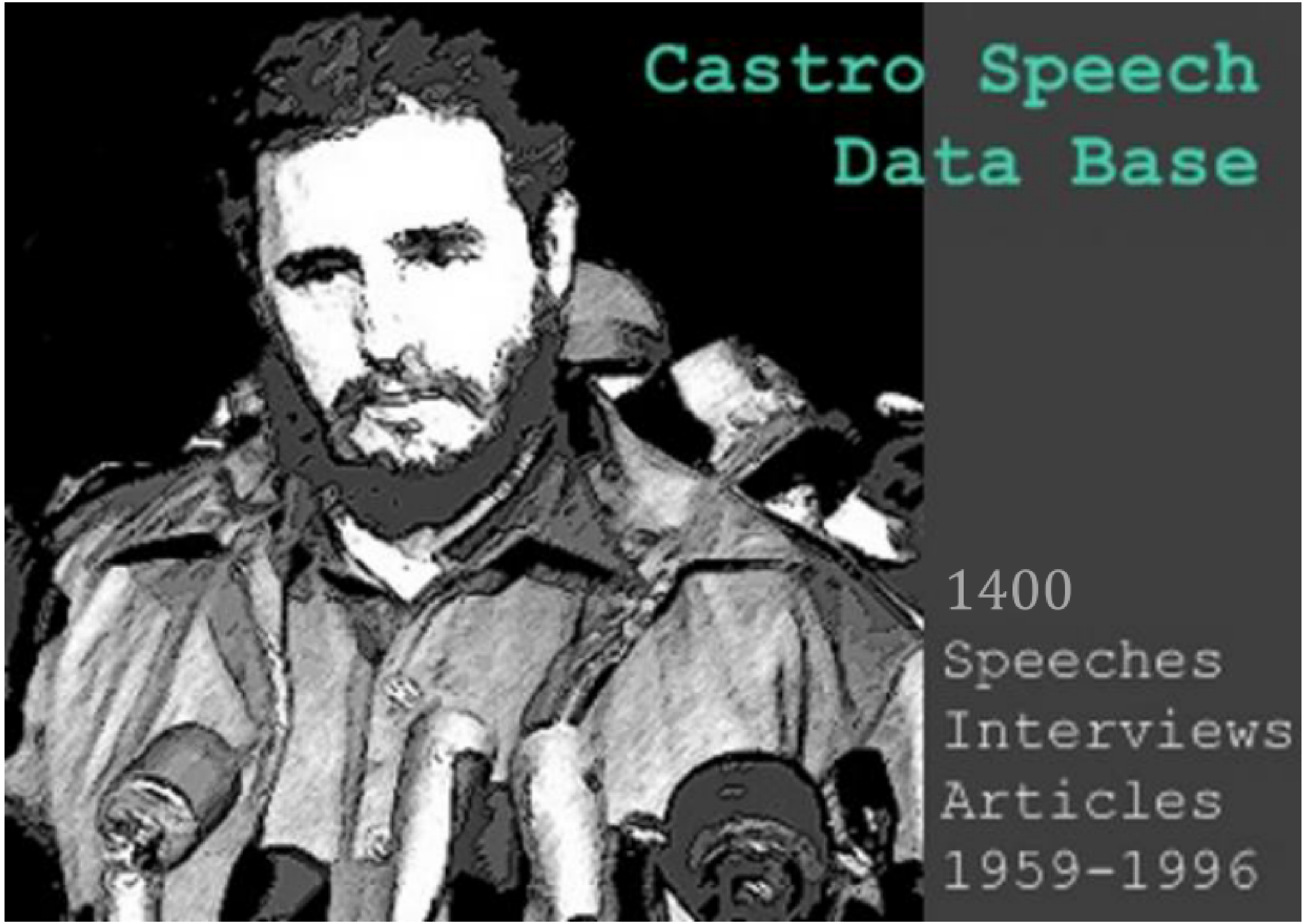
Document-Based Visualisation

GOAL

- Group together similar documents

How ?

- Historically relevant information is extracted
  - Persons, Locations, Organisations (**NER**)
- A document is represented by a vector of named entities
  - Similar documents have similar vectors (small cosine)
- A document collection is a graph
  - Nodes are documents, edges encode similarity
- The graph can be searched (integration in an IR system)


Castro Speech
Data Base
1400
Speeches
Interviews
Articles
1959-1996

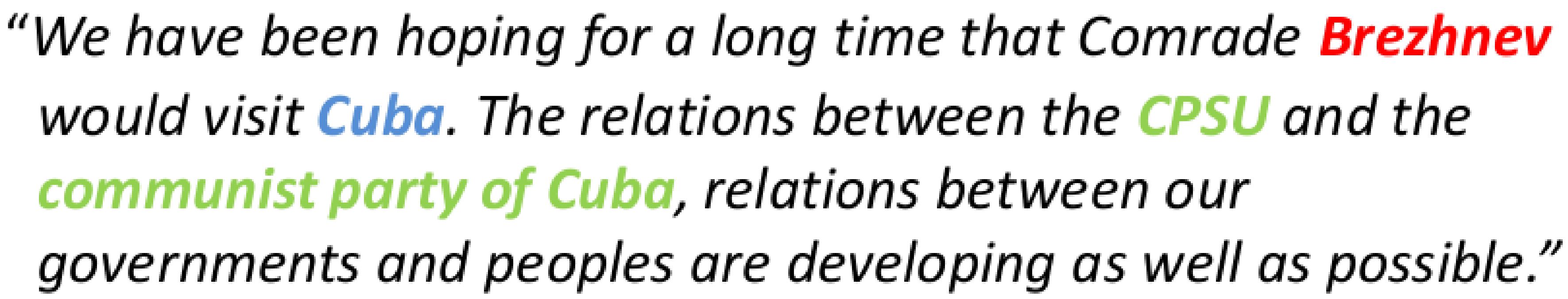
Named Entities

Persons

Locations

Organizations

Named Entity Recognition



*"We have been hoping for a long time that Comrade **Brezhnev** would visit **Cuba**. The relations between the **CPSU** and the **communist party of Cuba**, relations between our governments and peoples are developing as well as possible."*

Document Representation

$$\begin{array}{l}\text{Castro}\\ \text{Breznev}\\ \text{Batista}\\ \dots\end{array}\begin{bmatrix}1\\ 1\\ 0\\ \dots\end{bmatrix}\quad\begin{array}{l}\text{Cuba}\\ \text{Moscow}\\ \text{Volgograd}\\ \dots\end{array}\begin{bmatrix}1\\ 1\\ 1\\ \dots\end{bmatrix}\quad\begin{array}{l}\text{CPSU}\\ \text{Health Ministry}\\ \text{Columbian Army}\\ \dots\end{array}\begin{bmatrix}2\\ 3\\ 1\\ \dots\end{bmatrix}\quad\begin{array}{l}\text{declare}\\ \text{visit}\\ \text{relations}\\ \dots\end{array}\begin{bmatrix}6\\ 4\\ 3\\ \dots\end{bmatrix}$$

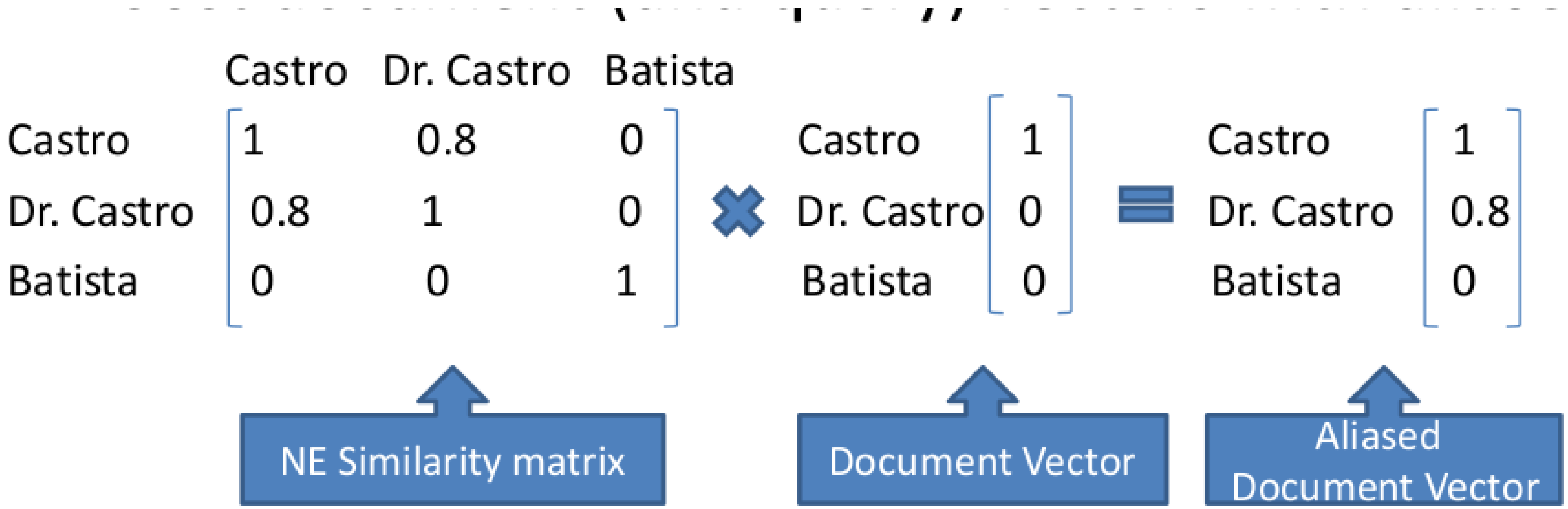
Alias Recognition

Fidel Castro
Dr. Fidel Castro
Maj. Fidel Castro
Castro
...

Cuban Communist Party
Communist Party of Cuba
...

- String Kernel to measure string similarity
- Helps recognizing linguistic variants of names in a collection of documents

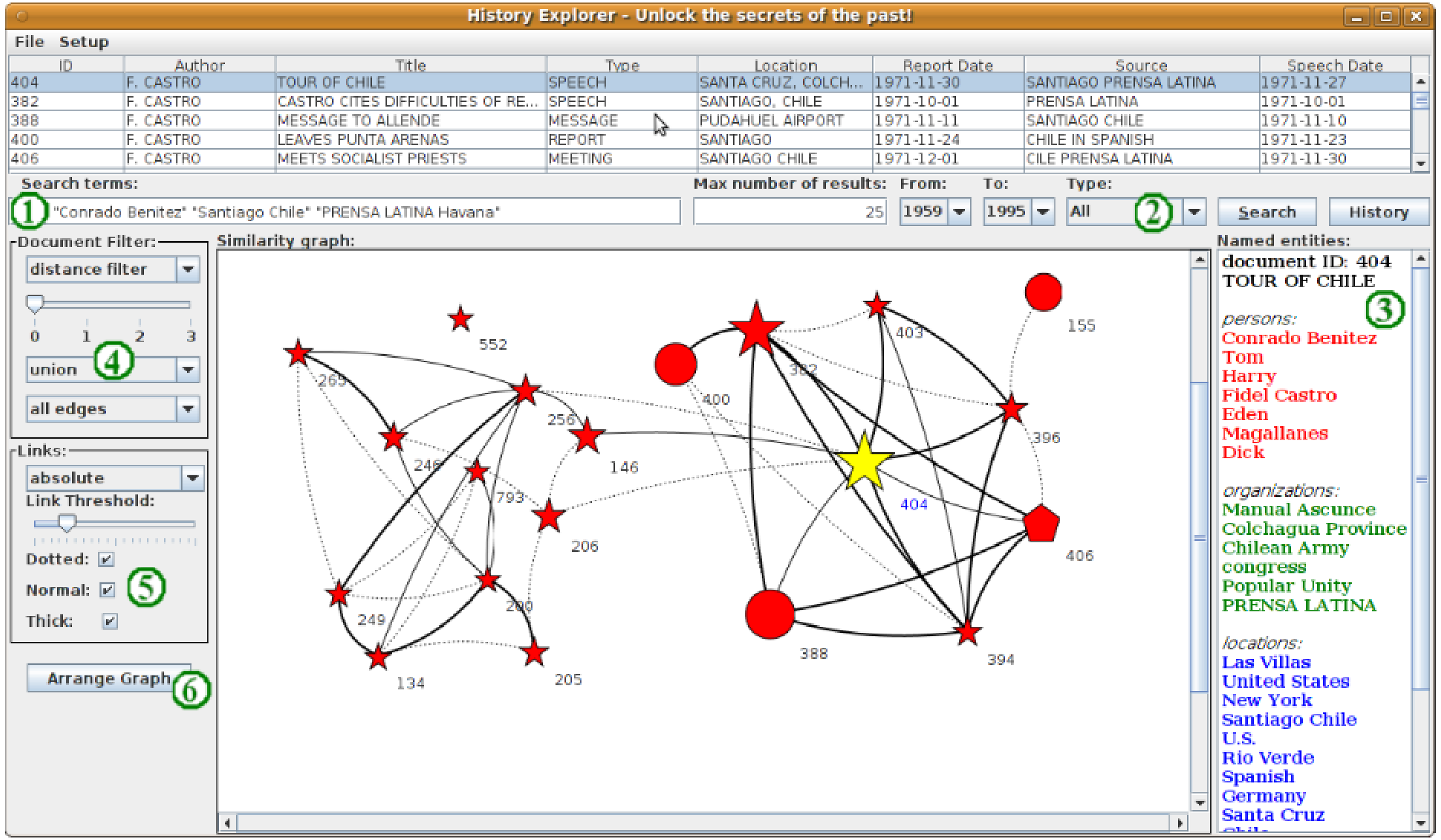
Alias Boosting



$$\begin{array}{c|ccc|} & \text{Castro} & \text{Dr. Castro} & \text{Batista} \\ \text{Castro} & 1 & 0.8 & 0 \\ \text{Dr. Castro} & 0.8 & 1 & 0 \\ \text{Batista} & 0 & 0 & 1 \end{array} \times \begin{array}{c|c|} \text{Castro} & 1 \\ \text{Dr. Castro} & 0 \\ \text{Batista} & 0 \end{array} = \begin{array}{c|c|} \text{Castro} & 1 \\ \text{Dr. Castro} & 0.8 \\ \text{Batista} & 0 \end{array}$$

- Aliases of names that appear in the documents receive an additional weight (more reliable document similarity measure)
- A similar expansion is used for queries that contain NEs (increased querying flexibility)

Graphical User Interface



History Explorer - Unlock the secrets of the past!

File Setup

ID	Author	Title	Type	Location	Report Date	Source	Speech Date
404	F. CASTRO	TOUR OF CHILE	SPEECH	SANTA CRUZ, COLCH...	1971-11-30	SANTIAGO PRENSA LATINA	1971-11-27
382	F. CASTRO	CASTRO CITES DIFFICULTIES OF RE...	SPEECH	SANTIAGO, CHILE	1971-10-01	PRENSA LATINA	1971-10-01
388	F. CASTRO	MESSAGE TO ALLENDE	MESSAGE	PUDAHUEL AIRPORT	1971-11-11	SANTIAGO CHILE	1971-11-10
400	F. CASTRO	LEAVES PUNTA ARENAS	REPORT	SANTIAGO	1971-11-24	CHILE IN SPANISH	1971-11-23
406	F. CASTRO	MEETS SOCIALIST PRIESTS	MEETING	SANTIAGO CHILE	1971-12-01	CILE PRENSA LATINA	1971-11-30

Search terms: "Conrado Benitez" "Santiago Chile" "PRENSA LATINA Havana"

Max number of results: 25

From: 1959

To: 1995

Type: All

Search History

Document Filter: distance filter

union

all edges

Links: absolute

Link Threshold:

Dotted: ☑

Normal: ☑

Thick: ☑

Arrange Graph

Similarity graph:

Named entities:

document ID: 404
TOUR OF CHILE

persons:
Conrado Benitez
Tom
Harry
Fidel Castro
Eden
Magallanes
Dick

organizations:
Manual Ascunce
Colchagua Province
Chilean Army
congress
Popular Unity
PRENSA LATINA

locations:
Las Villas
United States
New York
Santiago Chile
U.S.
Rio Verde
Spanish
Germany
Santa Cruz

- Query for keywords
- Results shown as interactive graph and as table

Benefits for Historians

Search + Visualisation

- Search: to express a specific historical question
- Visualisation: to help finding answers and formulating new questions

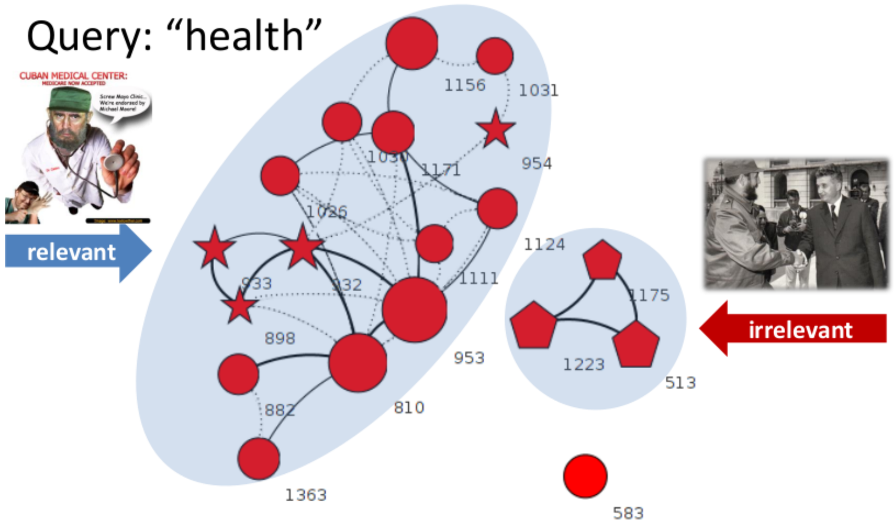
Helps ...

- Identify important topics (dense graph regions) and the corresponding documents
- Discover NEs that play an important role in a given topic
- Separating relevant from irrelevant documents

Clustering ...

- Can be used to further analyse similar documents

Relevant vs. Irrelevant



- Two groups of interconnected documents.
- The bigger group contains documents concerning health-care in Cuba in the 80s
- The smaller group is about greetings, wishing each other “good health”

Clustering

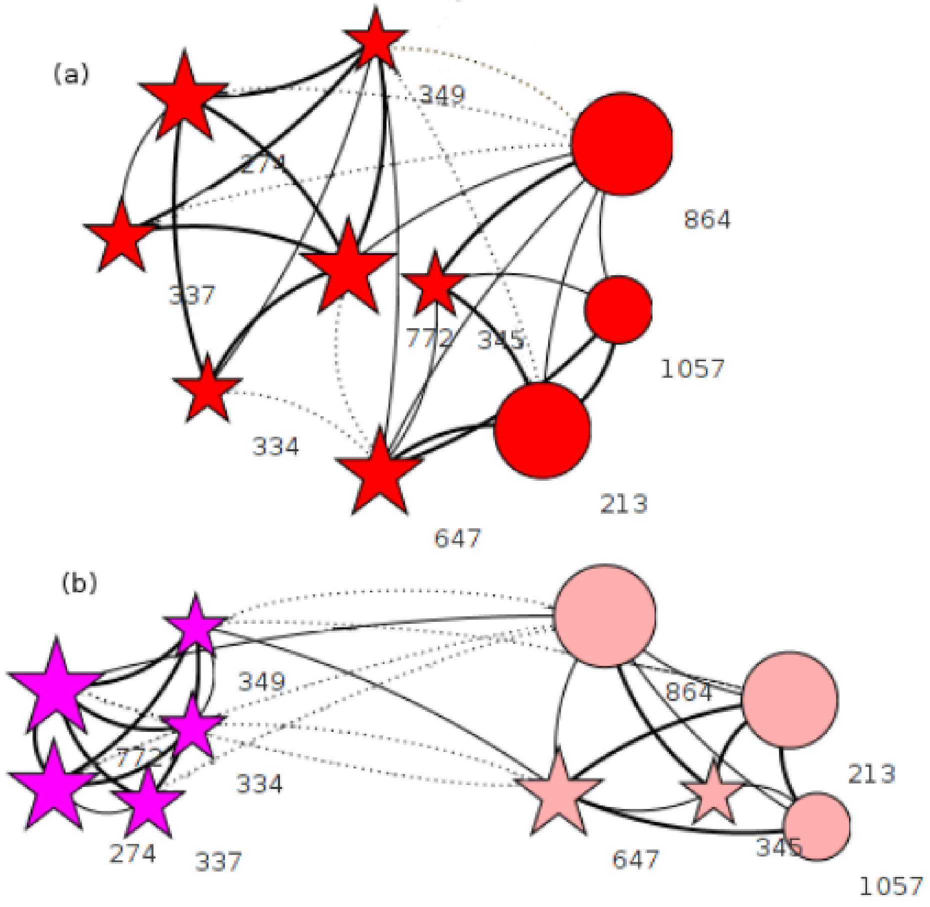


Figure 6. The resulting graphs for the query “Giron Kennedy” before clustering (a), and after clustering (b)

Clusters

- Documents directly related to the 1961 Bay of Pigs Invasion, e.g. speeches on the anniversary of the event or victory speeches
- Documents mentioning the invasion but not directly related to the event

NLP Tools used

Named Entity Recognition

- Stanford Named Entity Recognizer

Alias Recognition

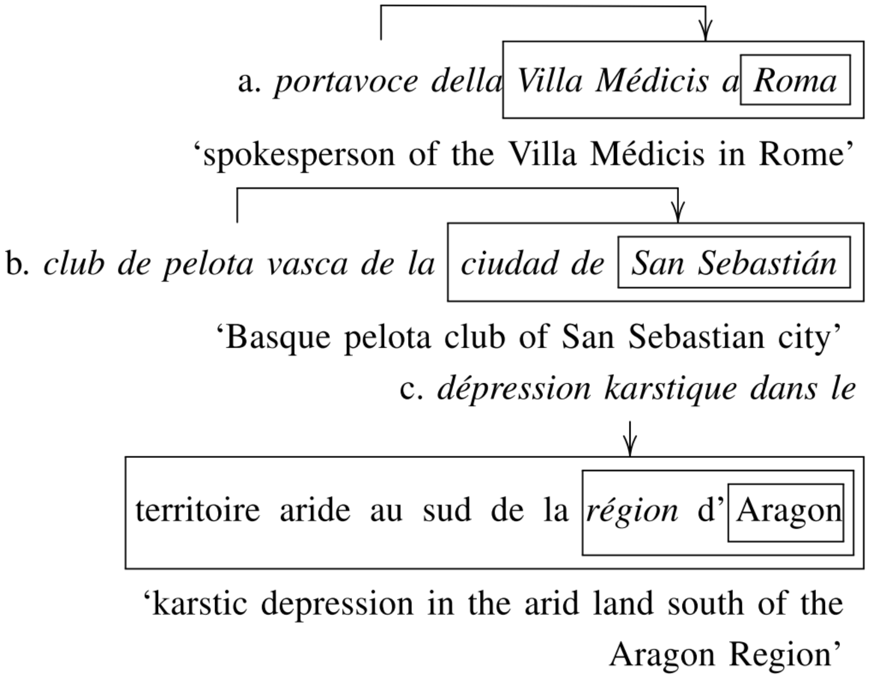
- String Kernel

Clustering

- Chinese Whispers Clustering Algorithm

Named Entity Recognition (NER)

Which mentions ?



a. *portavoce della Villa Médicis a Roma*

‘spokesperson of the Villa Médicis in Rome’

b. *club de pelota vasca de la ciudad de San Sebastián*

‘Basque pelota club of San Sebastian city’

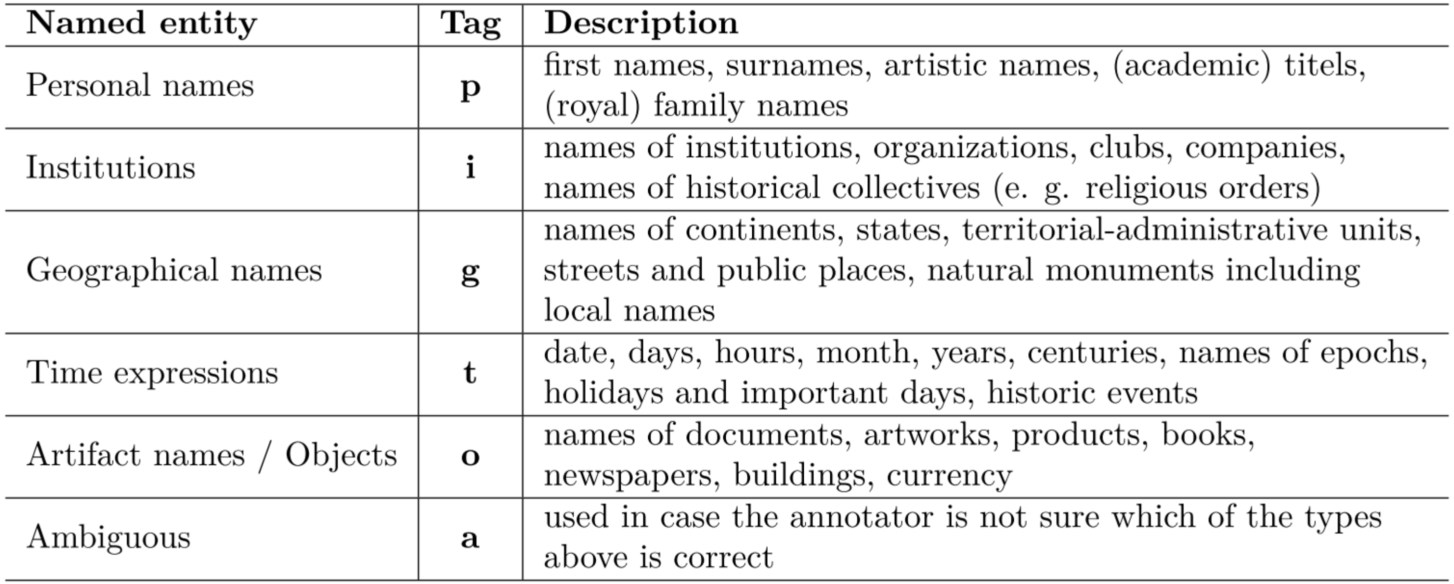
c. *dépression karstique dans le* territoire aride au sud de la *région* d’ Aragon

‘karstic depression in the arid land south of the Aragon Region’

Gaio and Moncla, IARA'17

Named Entity Recognition (NER)

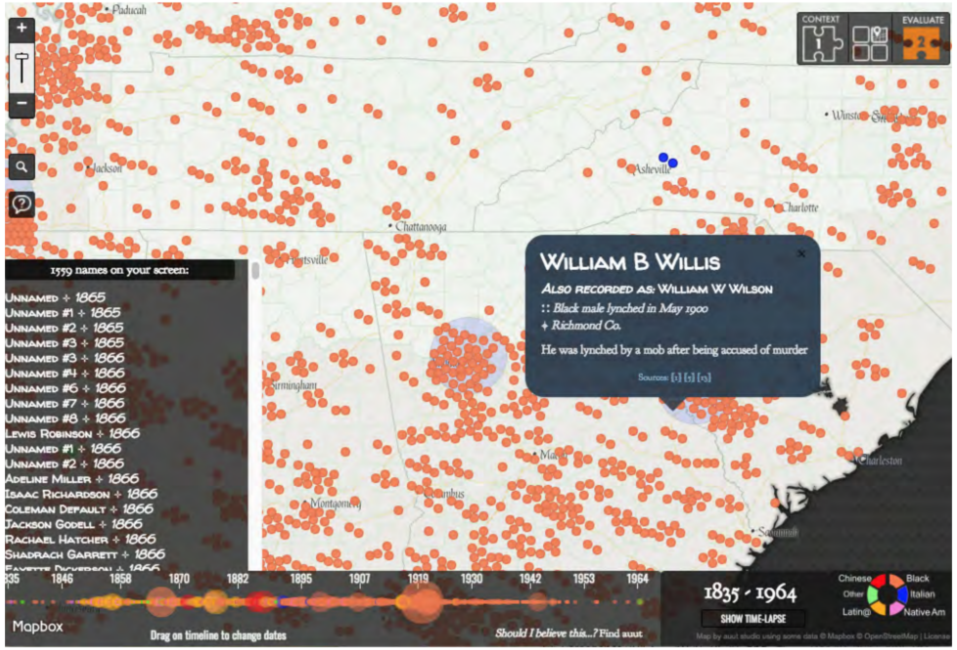
Which types ?



Named entity	Tag	Description
Personal names	**p**	first names, surnames, artistic names, (academic) titels, (royal) family names
Institutions	**i**	names of institutions, organizations, clubs, companies, names of historical collectives (e. g. religious orders)
Geographical names	**g**	names of continents, states, territorial-administrative units, streets and public places, natural monuments including local names
Time expressions	**t**	date, days, hours, month, years, centuries, names of epochs, holidays and important days, historic events
Artifact names / Objects	**o**	names of documents, artworks, products, books, newspapers, buildings, currency
Ambiguous	**a**	used in case the annotator is not sure which of the types above is correct

Czech Named Entity Corpus 2 . 0, Helena Hubkova, Pavel Kraal, Eva Pettersson LREC'18

Location-Based Visualisation



1559 names on your screen:

Unnamed ✦ 1865
Unnamed #1 ✦ 1865
Unnamed #2 ✦ 1865
Unnamed #3 ✦ 1865
Unnamed #3 ✦ 1866
Unnamed #4 ✦ 1866
Unnamed #6 ✦ 1866
Unnamed #7 ✦ 1866
Unnamed #8 ✦ 1866
Lewis Robinson ✦ 1866
Unnamed #1 ✦ 1866
Unnamed #2 ✦ 1866
Adeline Miller ✦ 1866
Isaac Richardson ✦ 1866
Coleman Default ✦ 1866
Jackson Godell ✦ 1866
Rachael Hatcher ✦ 1866
Shadrach Garrett ✦ 1866
Fayette Dickerson ✦ 1866

William B Willis

Also recorded as: William W Wilson

:: *Black male lynched in May 1900*

✦ *Richmond Co.*

He was lynched by a mob after being accused of murder

Sources: [1] [5] [15]

Goal

- Ground information to location

How ?

- Toponym Resolution: location names are resolved to a geographic reference

Recognising vs. Resolving Named Entities

Hamilton is in the North-Island

Recognising

- Named Entities

Hamilton, TYPE: LOC

- Aliases Recognition

Fidel Castro, Maj. Fidel Castro, Castro, Batista ...

- Corefering mentions

Hamilton is in the North Island.
It has a nice beach.
Weta is a big thing in that city

- Cross-document Coreferring Mentions

Resolving

- Entity Linking: NE ⇒ **Knowledge Base Entity**

Hamilton (Ontario)
Hamilton (New Zealand)
Hamilton (Ohio)
Hamilton (Bermuda)

- Geo-tagging: Toponym ⇒ **Geographic Reference**

Hamilton, Lat: -37.46, Lon: 175.16

Cross-Document Coreference Resolution

Document 1

UAW president Stephen Yokich then met separately for at least an hour with chief executives Robert Eaton of Chrysler Corp., Alex Trotman of Ford Motor Co. and finally with John Smith Jr. of General Motors Corp.

Document 2

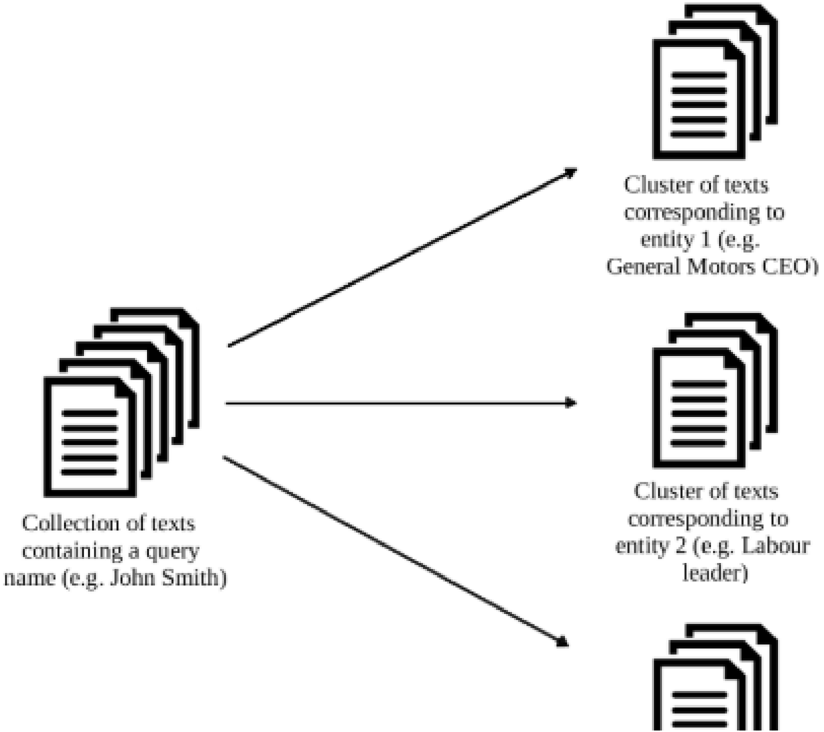
Blair became Labour leader after the sudden death of his successor John Smith in 1994 and since then has steadily purged the party of its high-spend and high-tax policies and its commitment to national ownership of industrial assets.

Document 3

Two years ago, Powell switched coaches from Randy Huntington to John Smith, who is renowned for his work with sprinters from 100 to 400 meters.

John Smith ⇒ 3 document clusters

- the CEO of General Motors
- the Labour Party leader
- an athletics coach


Collection of texts containing a query name (e.g. John Smith)
Cluster of texts corresponding to entity 1 (e.g. General Motors CEO)
Cluster of texts corresponding to entity 2 (e.g. Labour leader)

Why is NE Resolution hard ?

Language

Aliasing

- The same entity can have different names
- Location names often change over time

Ambiguity

- The same entity name can be used for different entities

Tools and Resources

Error Propagation

- NER errors propagate to NE disambiguation

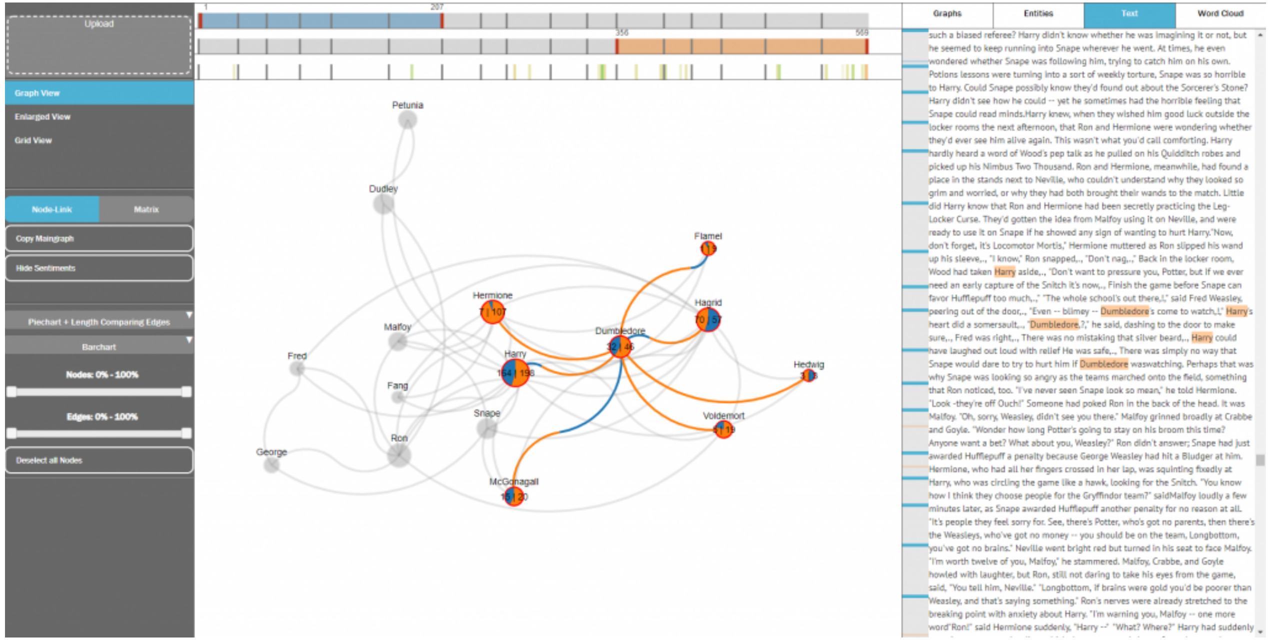
Domain Adpatation

- Most tools trained on contemporary Language

Incomplete or inadequate KB/Gazeteer

- The entity may not be present in the K/GazeteerB

Event-Based Visualisation



such a biased referee? Harry didn't know whether he was imagining it or not, but he seemed to keep running into Snape wherever he went. At times, he even wondered whether Snape was following him, trying to catch him on his own. Potions lessons were turning into a sort of weekly torture, Snape was so horrible to Harry. Could Snape possibly know they'd found out about the Sorcerer's Stone? Harry didn't see how he could -- yet he sometimes had the horrible feeling that Snape could read minds.Harry knew, when they wished him good luck outside the locker rooms the next afternoon, that Ron and Hermione were wondering whether they'd ever see him alive again. This wasn't what you'd call comforting. Harry hardly heard a word of Wood's pep talk as he pulled on his Quidditch robes and picked up his Nimbus Two Thousand. Ron and Hermione, meanwhile, had found a place in the stands next to Neville, who couldn't understand why they looked so grim and worried, or why they had both brought their wands to the match. Little did Harry know that Ron and Hermione had been secretly practicing the Leg-Locker Curse. They'd gotten the idea from Malfoy using it on Neville, and were ready to use it on Snape if he showed any sign of wanting to hurt Harry."Now, don't forget, it's Locomotor Mortis," Hermione muttered as Ron slipped his wand up his sleeve,., "I know," Ron snapped,., "Don't nag,.," Back in the locker room, Wood had taken Harry aside,., "Don't want to pressure you, Potter, but if we ever need an early capture of the Snitch it's now,., Finish the game before Snape can favor Hufflepuff too much,.," "The whole school's out there,!," said Fred Weasley, peering out of the door,., "Even -- blimey -- Dumbledore's come to watch,!," Harry's heart did a somersault,., "Dumbledore,?," he said, dashing to the door to make sure,., Fred was right,., There was no mistaking that silver beard,., Harry could have laughed out loud with relief He was safe,., There was simply no way that Snape would dare to try to hurt him if Dumbledore waswatching. Perhaps that was why Snape was looking so angry as the teams marched onto the field, something that Ron noticed, too. "I've never seen Snape look so mean," he told Hermione. "Look -they're off Ouch!" Someone had poked Ron in the back of the head. It was Malfoy. "Oh, sorry, Weasley, didn't see you there." Malfoy grinned broadly at Crabbe and Goyle. "Wonder how long Potter's going to stay on his broom this time? Anyone want a bet? What about you, Weasley?" Ron didn't answer; Snape had just awarded Hufflepuff a penalty because George Weasley had hit a Bludger at him. Hermione, who had all her fingers crossed in her lap, was squinting fixedly at Harry, who was circling the game like a hawk, looking for the Snitch. "You know how I think they choose people for the Gryffindor team?" saidMalfoy loudly a few minutes later, as Snape awarded Hufflepuff another penalty for no reason at all. "It's people they feel sorry for. See, there's Potter, who's got no parents, then there's the Weasleys, who've got no money -- you should be on the team, Longbottom, you've got no brains." Neville went bright red but turned in his seat to face Malfoy. "I'm worth twelve of you, Malfoy," he stammered. Malfoy, Crabbe, and Goyle howled with laughter, but Ron, still not daring to take his eyes from the game, said, "You tell him, Neville." "Longbottom, if brains were gold you'd be poorer than Weasley, and that's saying something." Ron's nerves were already stretched to the breaking point with anxiety about Harry. "I'm warning you, Malfoy -- one more word"Ron!" said Hermione suddenly, "Harry --" "What? Where?" Harry had suddenly

Goal

- Identify entity networks (social networks)

How ?

- NER and Text Segmentation
- Entities that cooccur within a segment are groupe withing a network

Social Network Visualisation

Interactive, semi-automatic Expert-Driven Data Analysis

- Automatic tools are imperfect and their results can lead e.g., to highly skewed impressions of the relative importance of characters in a text

- Interaction with domain experts helps mitigate these issues and support the development of better, more adapted NLP tools

Generic Workflow

The workflow has been applied to

- narrative (modern and medieval) texts
- theoretical philosophical texts, with the goal of establishing relations between philosophical networks
- parliamentary debates, with the goal of connecting po- litical parties to political issues

Blessing, Echelmeyer, John and Reiter. 2017

Workflow

Text Segmentation

NE linking

Create networks of entities that co-occur within a segment

- E.g., the characters that take part in a great feast

Manual exploration of networks for validation

Semi-Automatic Processing

Automatic pre-processing

- tokenization, sentence segmentation, POS tagging

NER

- Manual annotation of NEs (persons and locations)
- Training a NER (language specific gazetteers and POS tagging for the features)
- NER evaluation using cross validation
- Additional manual annotations (to improve recall)

NE Resolution

- Manual entity grounding (to a pre-defined list of characters)

Semi-Automatic Processing (Ct'd)

Segmentation

- Automatic segmentation (sentence, paragraphs)

Web-Based Exploration

- Web-based tool for close and distant reading
    - User can view the text passages of the selected entities
    - Network graphs are created with Gephi (Bastian et al., 2009), which provides various layout algorithms, offers statistics and network metrics

Questions

Which visualisation method ?

- Explore vs. Display
- One or several

Named Entities

- Recognition or Disambiguation ?
- Mentions and Types ?

Domain Adaptation

- Language: Contemporary → Historical (Text Normalisation vs. Tool Adpatation)
- Domain: National Newspaper, blogs → Regional journal, Encyclopedia

Thanks