Having fun in Paris, New York!

You mean Paris, France! The one with the Eiffel Tower?

Spatial Entity Disambiguation in Social Media Content

Amosse Edouard, PhD Student
Nhan LE-THAN, Supervisor

Outlines

- Introduction
  - Context
  - Motivation & Research Question
- Overview of the State of the art
  - Location prediction on social media content
- Named Entity Recognition Process
  - Recognition
  - Disambiguation
  - Identification
- The proposed approach
  - Spatial Entity Modeling
  - Spatial Entity Disambiguation and Identification
- Conclusion And Perspectives

The Web Today

- We have moved a huge part of our social life online.
- Numerous tools
  - Blogs, Microblogs,
  - Social networks
  - Sensors network
- Freedom of the crowd
  - We see, we hear, we thing → we share

The Need ...

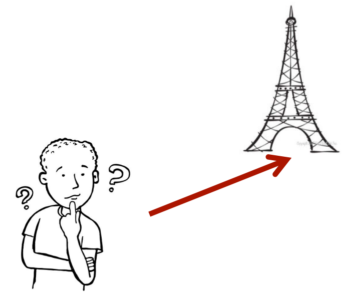
- To go from Data → Information
- From Information → Information meaning
- Enable machines to understand data on the Web

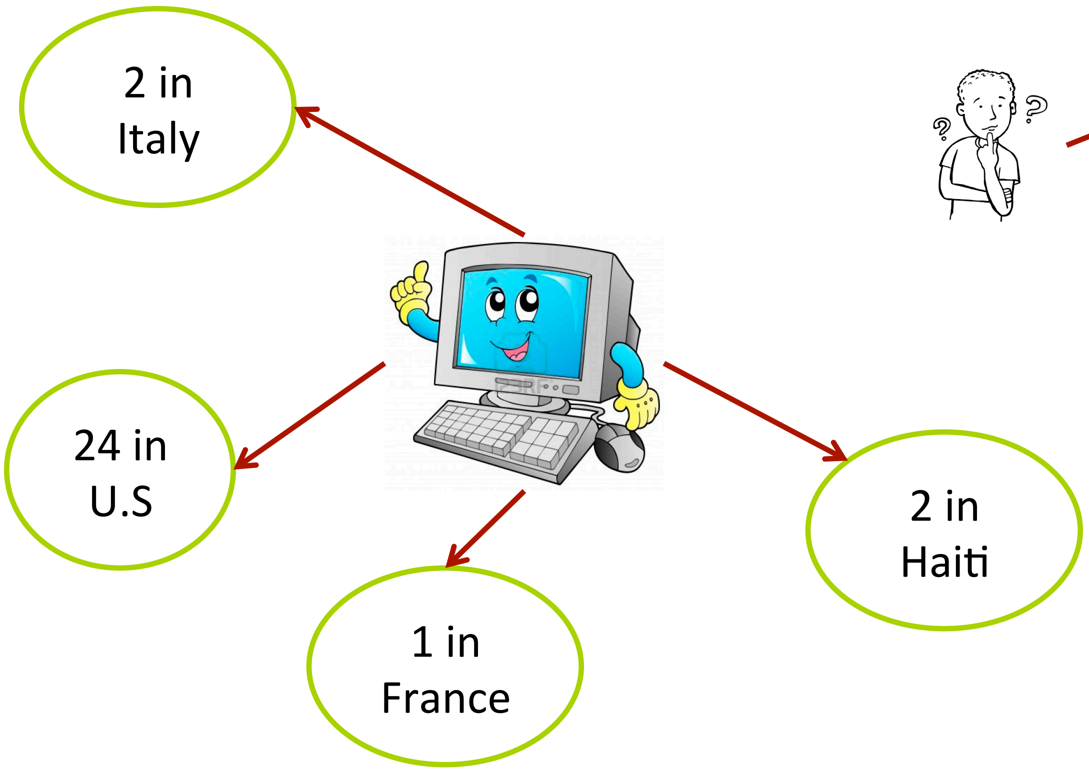
Social Media Content

- Context dependent
- Limited in space in time
- Lack of structure

Context

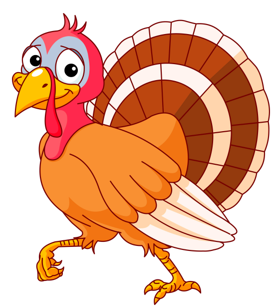
I am having fun in Paris.





Motivation

- **Most spatial entity are ambiguous [4]**
  - **Paris (France, US), New York (US, South Africa**
- **Spatial entities might have different meanings**



Research Questions

Predicting location of social media content

- **Spatial Entity Identification**
- **Spatial Entity Disambiguation**

Why we Care?

- Less than 1% of twitter posts are geo located [1]
- A content might have many locations references
  - User location
  - Event location
  - Server location
- Spatio-temporal analysis of social media
  - Information filtering
  - User trend by location
  - Event detection

Who Cares?

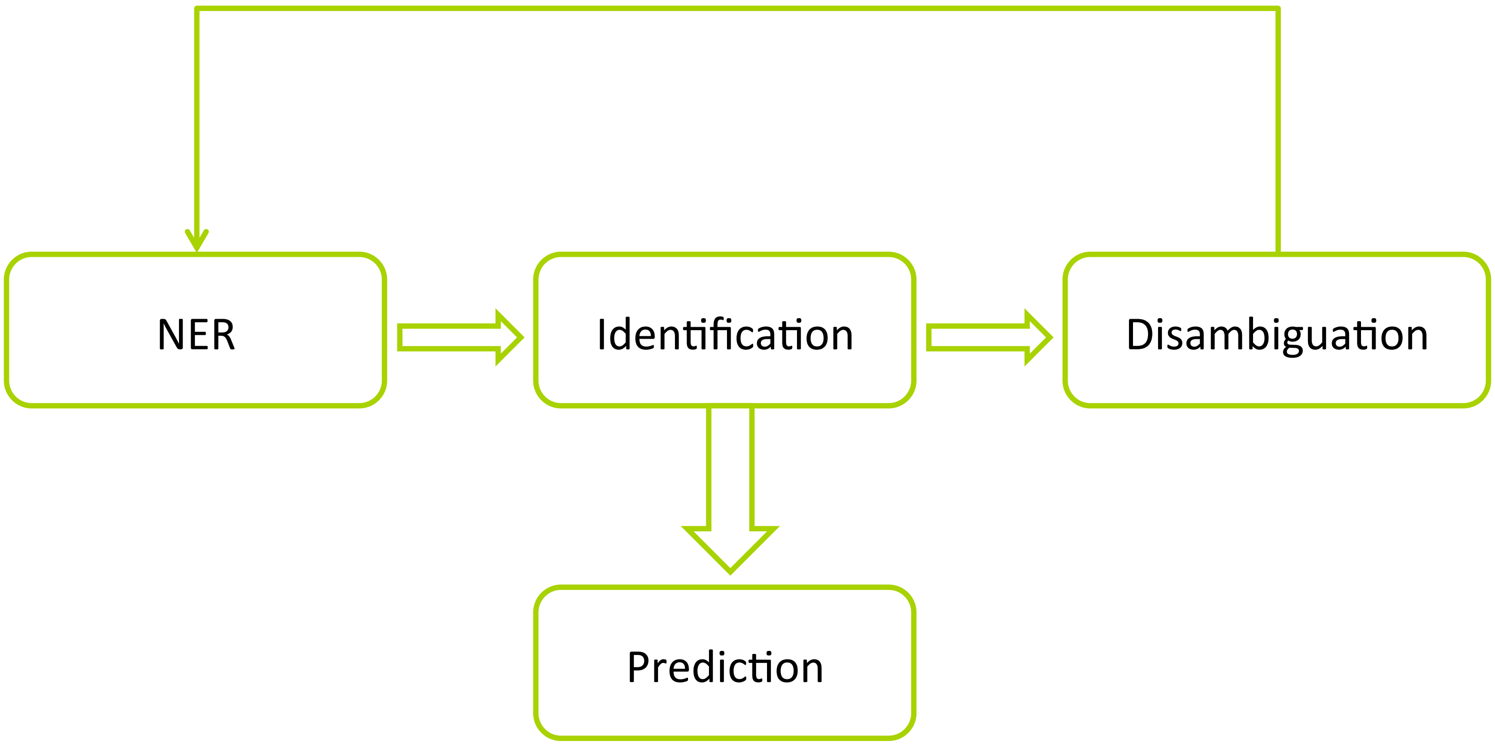
Location prediction on social media content

- Naive Bays approach by S. Kinsella [1]
- Geolocated Flickr photos [2]
- DBPedia Spotlight [3]

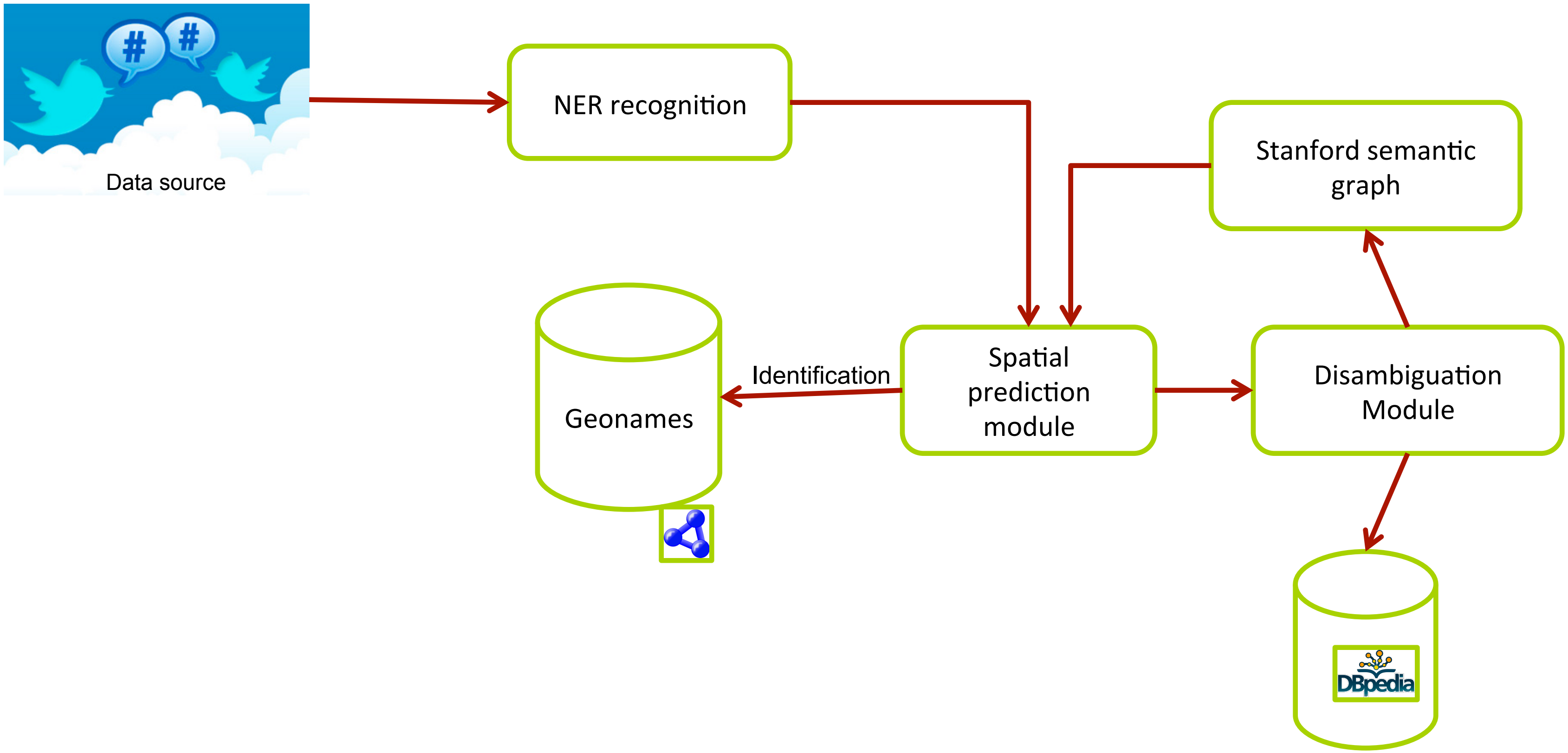
Web pages

- Web a Where Geotagging Web Content [4]

Location Prediction Process

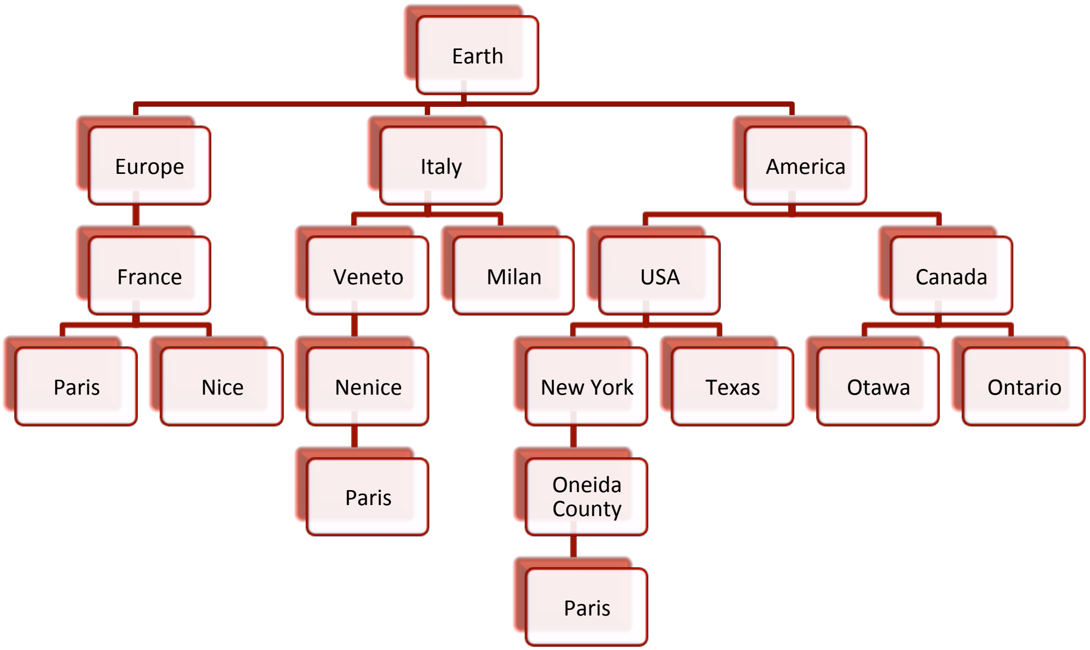


The Proposed Approach



Spatial Entity Modeling

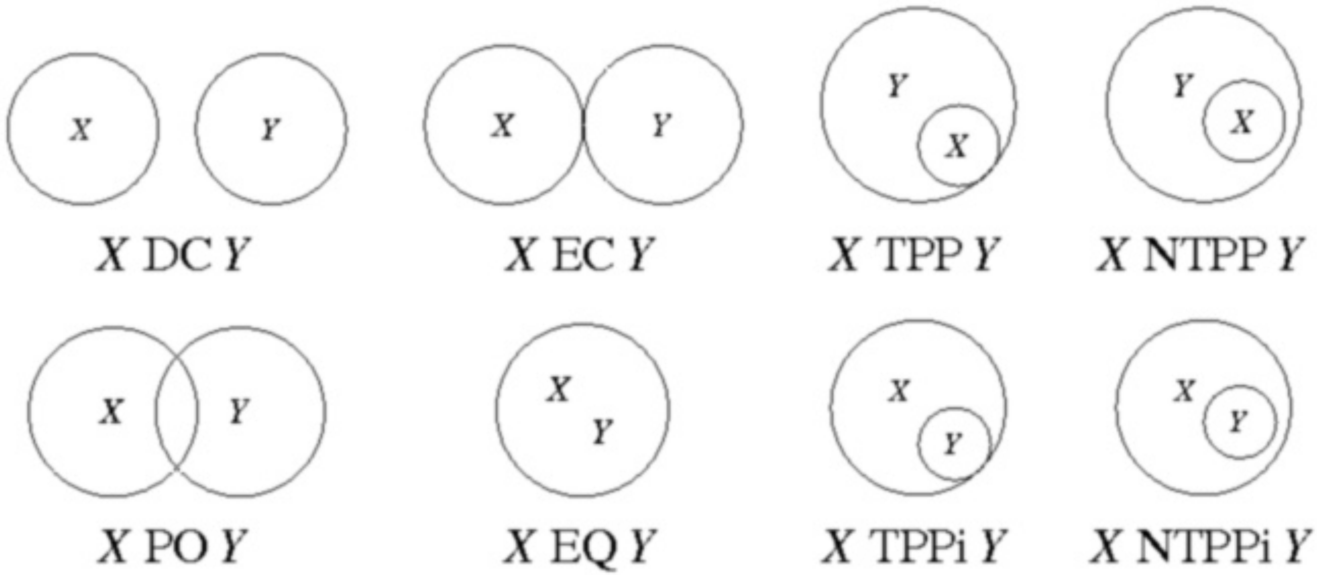
- The world is organized such as a hierarchical tree



Spatial Entity Topological Relation

The Region Connection Calculus [5]

- RCC-8

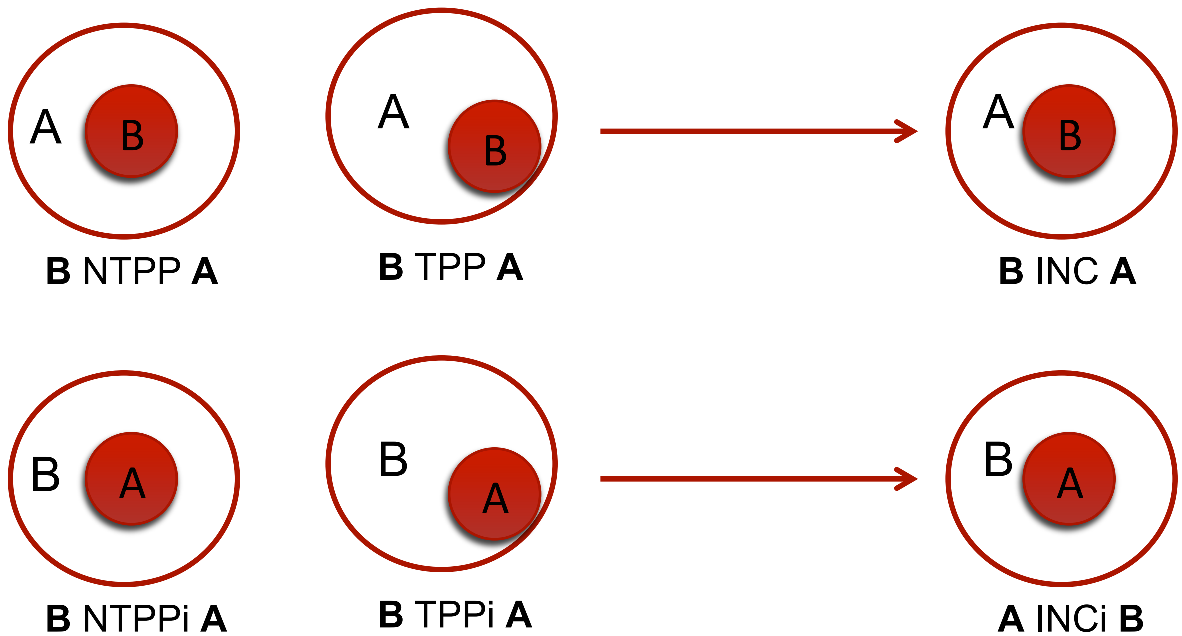


RCC-8 → RCC-5

- RCC-8 → RCC-5
  - DJ : A is **disjoint** with B
  - INC: A is **included** in B
  - INCi: A **contains** B
  - EQ: A is **equal** to B
  - PO : A **partially overlaps** B

RCC-8	RCC-4
A DC B – A EC B	A DJ B
A TPP B – A NTPP B	A INC B
A TPPi B – A NTPPi B	A INCi B
A EQ B	A EQ B
A PO B	A PO B

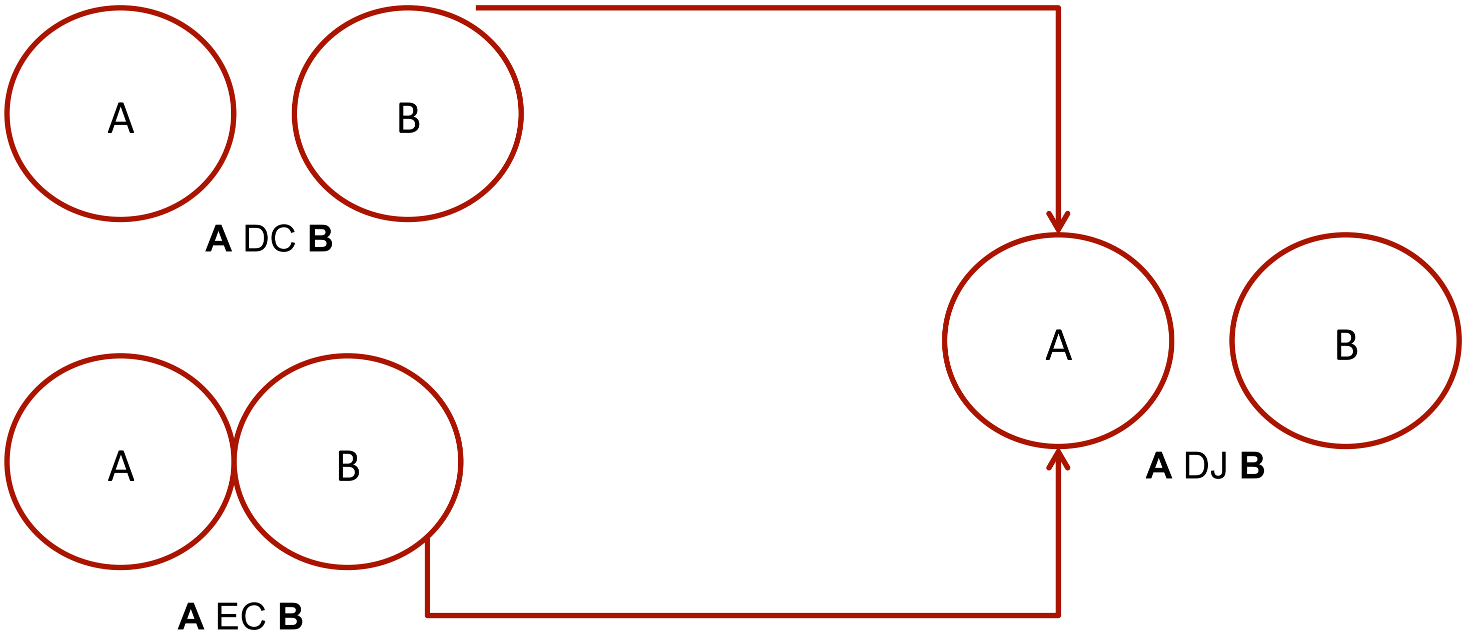
Inclusion Relation



$$INC(B,A) = B \subseteq A \leftrightarrow \forall\, p \subset P(B)\,,\, p \subset P(A)$$

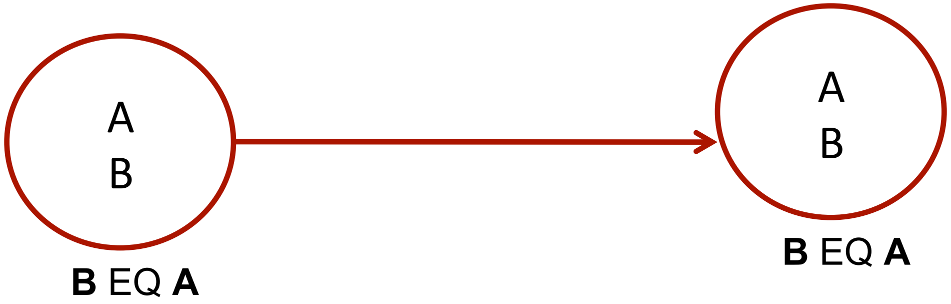
$$INCi\ (B,A) = \exists\, p \subset P(A)\ /\ p{=}B$$

Disjoint Relation



$$DJ\ (B,A) \leftrightarrow A \cap B = \Theta \rightarrow \{\nexists\ p \subseteq P(A) / p \subseteq P(B)\}$$

Equality Relation



$$EQ\ (B,A) \leftrightarrow (A=B) = INC(A,B) \land INC(B,A)$$

Partial Overlaping

B PO **A** **B** PO **A**

$$PO\ (B,A) \leftrightarrow (A \cap B \neq \Theta \wedge (A \cap B \neq A \wedge B \cap A \neq A))$$
$$= \{\exists\, p \,/\, p \subseteq P(A) \wedge p \subseteq P(B)\}$$

Named Entity Recognition

- Spotting a text to find named entities
- Common named entities
  - Person
  - Organisation
  - Place

Spatial Entity Recognition

- We use the Stanford NER parser
  - Input : Text
  - Output : Named entities
- No spatial Entity found
  - Enrich the content with data from URL if present
  - Repeat the process

Spatial Relation with Regular Expression

- Entities separated by a comma imply inclusion
  - Paris, New York → INC (Paris, New York)
- Entities separated by a coordinating conjunction are disjoint
  - Paris and New York → DJ (Paris, New York)
  - Paris, New York, Lisborn …
    - DJ (Paris, New York) – DJ (New York, Lisborn)

The Gazetteer

- We have built a local gazetteer with free data obtained from Geonames
  - 114 330 462 RDF triples
  - 7852909 spatial features
- Data are stored in RDF format

Spatial Entity Identification

- The local gazetteer is used
- Sparql-Query the local gazetteer
  - All feature within the same name
  - Linked to their feature parents
- The result is a hierarchical tree

Use of Spatial Relation

- Hypothesis
  - Single sense per discourse
  - Entities appeared within the same context are considered related
- We first determine the relation between entities
- We then build appropriate SPARQL Query to build the hierarchy

Disjoint Query

```
Select
        ?s
        (SAMPLE (?nn) as ?name)
        (SAMPLE (?code) as ?ccode)
        (SAMPLE (?dpPedia) as ?dpPediaUri)
        (SAMPLE (?clazz) as ?class)
        (GROUP_CONCAT(distinct ?cparent; separator="->") as ?parents)
        where
        {
        ?s gn:parentFeature+ ?parent.
        ?parent gn:name ?pname.
        ?s gn:featureClass ?clazz.
        ?s gn:featureCode ?code.
        ?s gn:name ?nn.
        BIND(CONCAT(?parent, ";;", ?pname) AS ?cparent).
        OPTIONAL
        {
        ?s rdfs:seeAlso ?dpPedia
        }
          {
           ?s gn:name "Paris".
          }
        }
GROUP BY ?s
```

Inclusion Query

```
Select
        ?s
        (SAMPLE (?nn) as ?name)
        (SAMPLE (?code) as ?ccode)
        (SAMPLE (?dpPedia) as ?dpPediaUri)
        (SAMPLE (?clazz) as ?class)
        (GROUP_CONCAT(distinct ?cparent; separator="->") as ?parents)
        where
        {
        ?s gn:parentFeature+ ?parent.
        ?parent gn:name ?pname.
        ?s gn:featureClass ?clazz.
        ?s gn:featureCode ?code.
        ?s gn:name ?nn.
        BIND(CONCAT(?parent, ";;", ?pname) AS ?cparent).
        OPTIONAL
        {
        ?s rdfs:seeAlso ?dpPedia
        }
        ?s gn:parentFeature+ ?parent_1.
        ?s gn:name "Paris".
        {
        select distinct ?parent_1 where
        {
         ?parent_1 gn:name "New York".
        }
        }
GROUP BY ?s
```

Thanks to O. Corby!

Spatial Entity Disambiguation

- Enrich ambiguous spatial entity
- Based on
  - Natural Language Processing
  - Linked Data

DBPedia Disambiguation

- We query DBPedia for additional information about the entity:
  - Description
  - Geo Location
  - A set of spatial-related properties

Disambiguation rule

- Spatial inclusion
  - A dbp:isPartOf B → INC (A,B)
  - A dbp:capital B → INC (B,A)
  - A dbp:part B → INC(B,A)
- Syntactic analysis
  - Stanford dependency graph
  - Qakis Relation Pattern

Syntactic Analysis – DBPedia Resources

- Text from DBPedia are well edited
- Syntactic analyzer will perform better than on short text
- Description of spatial entity often refer to related spatial entities

Qakis Relation Patterns

- Link named entities with dbpedia properties by applying NLP analysis
- We build a set of dbpedia properties that describe spatial inclusion
  - isPartOf, capital, country ….
- Retain from the result properties that are in the list and above a threshold

Qakis - Relation Patterns

- Paris is the capital and most populous city of France.
- [LOCATION] is the capital and most populous city of [LOCATION].

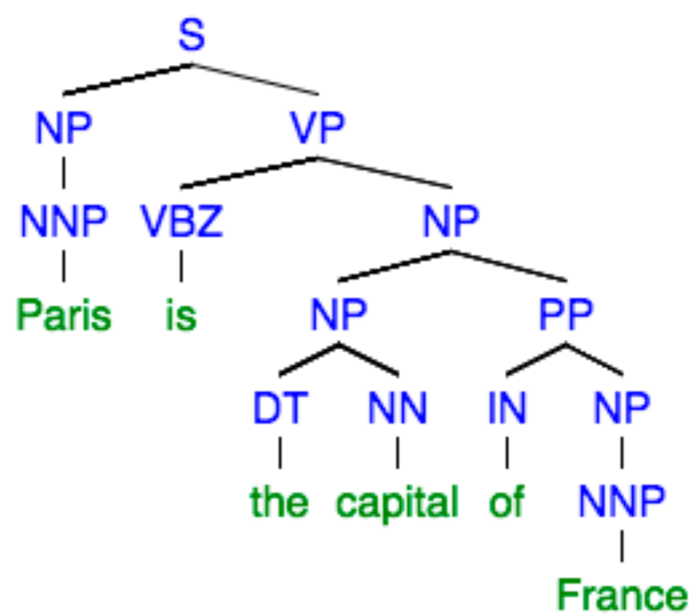
dbo:isPartOf;9.5610166666667;
dbo:twinCity;7;
dbo:largestCity;4.7920833333333;
dbo:capital;4.6932333333333;
dbo:city;4.5034833333333;
dbo:part;4.3656333333333;

Paris isPartOf France INC(Paris, France)
Paris capital France INC(Paris, France)

Stanford Semantic Tree

- Syntactic dependency graph
- Words are nodes and grammatical relation are edge

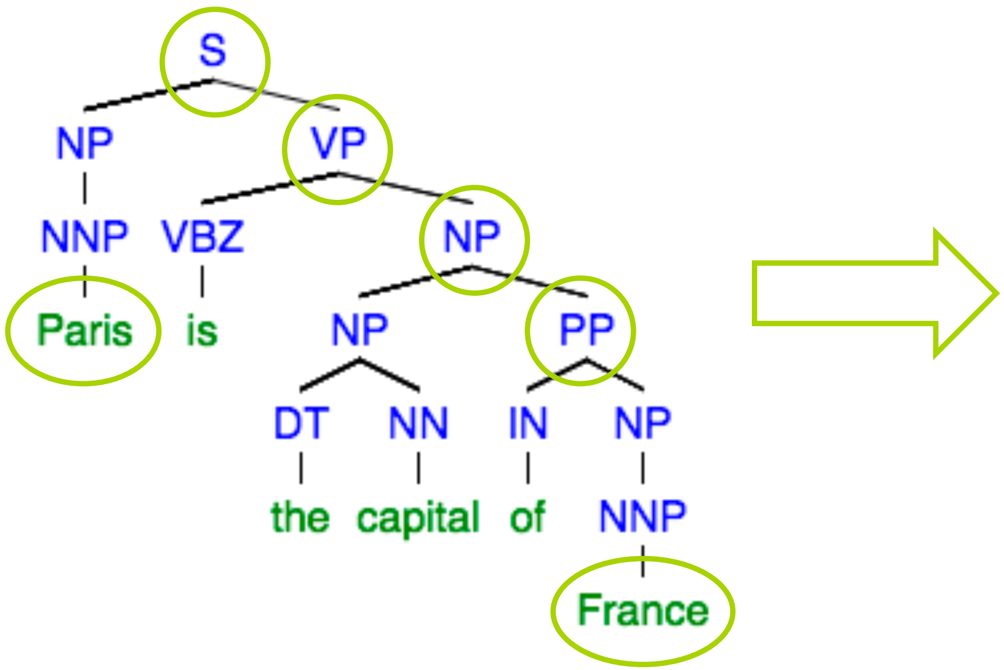
root (ROOT-0 , capital-4)
nsubj (capital-4 , Paris-1)
cop (capital-4 , is-2)
det (capital-4 , the-3)
prep_of (capital-4 , France-6)



Analyzing the Dependency Graph

- We have built a set of prepositions related to location
- Build spatial entity relation
  - We navigate into the graph started from the subject node
- Two entities are spatially related :
  - They have a common subject node
  - The path contains a spatial-related preposition

Analyzing the Stanford Graph



Paris, France
INC (Paris, France)

Discussion

- Place name may refer non spatial entities
  - *Washington is visiting Paris*
  - Obama granted pardon to a turkey
- In the future, we will consider all possible meanings of a term
- We may also apply semantic POS analysis
  - An entity followed by an action verb is more likely a Person than a Spatial thing.

Really? What about this "Paris has woken up under snow today"

Conclusion and Perspectives

- We propose an approach for spatial entity disambiguation based on
  - Natural Language Processing
  - Linked Data
- We are currently limited ourselves to English language
- The semantic dependency graph is a work ongoing
- Evaluation of the approach on a twitter dataset

References

1. M. Kinsella, Geolocation using Language Models from Geotags
2. Pavel Serdyukov, Placing Flickr Photos on a Map
3. J. Daiber et al., Improving Efficiency and Accuracy in Multilingual Entity Extraction
4. E. Amitay, Web-a-Where: Geotagging Web Content
5. D. Randal et al., A Spatial Logic based on Regions and Connection

Question and Suggestion

Thank you for your attention.