INFORMATION EXTRACTION IN METADATA: CHALLENGES AND TOOLS

Nuno Freire
The European Library
nfreire@gmail.com

Europeana Tech Conference

Vienna, October 2011

Outline

- Introduction to Information Extraction
  - Named entity recognition
  - Entity resolution
- Challenges of Information Extraction in metadata records
- Examples of available tools
- The EuropeanaConnect Geoparser
- Conclusion

Information extraction

- Structuring and combining data that is found, explicitly stated or implied, in text.
- The final output consists in semantically clean data
  - Following a well defined data model
  - On which computation can be applied
- Mainly supported by natural language processing techniques
  - Text tokenization, part-of-speech tagging, gramatical analysis, etc.

Named entity recognition

- Named entity recognition
  - Concerns the delimiting, in unstructured text, of the character strings that refer to entity names, and the classification of the entity type.
- Extensively researched in the area of natural language processing
  - text tokenization, part-of-speech tagging, word sequence analysis, etc.
- References to entities are recognized by reasoning on evidence provided by natural language.
  - State of the art techniques achieve near human performance
- Entity names are very relevant for information retrieval
  - 70% of queries in internet search engines contain named entities

Named entity recognition + entity resolution

- Entity recognition or disambiguation
  - Concerns the resolution of a name to a specific real-world entity.

- Entity resolution is becoming more relevant with the emergence of Linked Open Data
  - Comprehensive data sets are available to link entities to
    - VIAF – persons and organizations
    - Geonames – places
    - Dbpedia - covers many entity types

Information extraction in metadata records

- Although state of the art entity recognition achieves near human performance, when applied to metadata records, it underperforms.
  - These techniques are based on natural language processing, which requires well-structured text
  - Metadata fields often contain just short sentences or simple expressions
    - Lack of lexical evidence
- Particular challenges in Europeana metadata
  - Very large number of languages
  - Natural language processing tools not available for all languages

Examples

<record>
− <dc:title>
The Great South ;; a record of journeys in Louisiana, Texas, the Indian Territory, Missouri, Arkansas, Mississippi, Alabama, Georgia, Florida, South Carolina, North Carolina, Kentucky, Tennessee, Virginia, West
</dc:title>
<dc:creator>King, Edward</dc:creator>
− <dcterms:tableOfContents>
To Mr. Roswell- Smith; Louisiana, Past and Present; The French Quarter of New Orleans - The Revolution and its Effects; The Carnival - The French Markets; The Cotton Trade - The New Orleans Levées; The Canals and the Lake - The American Quarter; On the Mississippi River - The Levée System - Railroads - The Fort St. Philip Canal; The Industiries of Louisiana - A Sugar Plantation - The Teche Country; The Political Situation in Louisiana; Ho! for Texas - Galveston; A Visit to Houston; Pictures from Prison and Field; Austin, the Texan Capital - Politics - Schools; The Truth About Texas - The Journey by Stage to San Antonio; Among the Old Spanish Missions; The Pearl of the South-west; The Plains - The Cattle Trade; Denison - Texan Characteristics; -[...]
</dcterms:tableOfContents>
</record>

Examples

```
<record>
  <dc:title>Besuch bei der alten Dame</dc:title>
  <dc:creator>Rox-Schulz, Heinz</dc:creator>
- <dcterms:spatial>
    Ort: Peru - Occucaje - Cerro Córdoba (Herstellungsort)
  </dcterms:spatial>
- <dc:description>
    Das s/w-Foto zeigt Rox bei Grabungen im Ica-Tal in Peru neben zwei
    Mumien. Die linke Mumie brachte er mit nach Deutschland. Sie ist eines
    der Highlights der Sammlung des Abenteuermuseums Saarbrücken.
  </dc:description>
</record>
```

Some available tools

- GATE - http://gate.ac.uk/
  - Provides several information extraction functionalities
  - Supports several languages
- Stanford NER
  - Named entity recognition
  - English - http://nlp.stanford.edu/ner/
  - German - http://www.nlpado.de/~sebastian/software/ner_german.shtml
- OpenNLP - http://incubator.apache.org/projects/opennlp.html
  - Named entity recognition
  - English, Spanish, Dutch

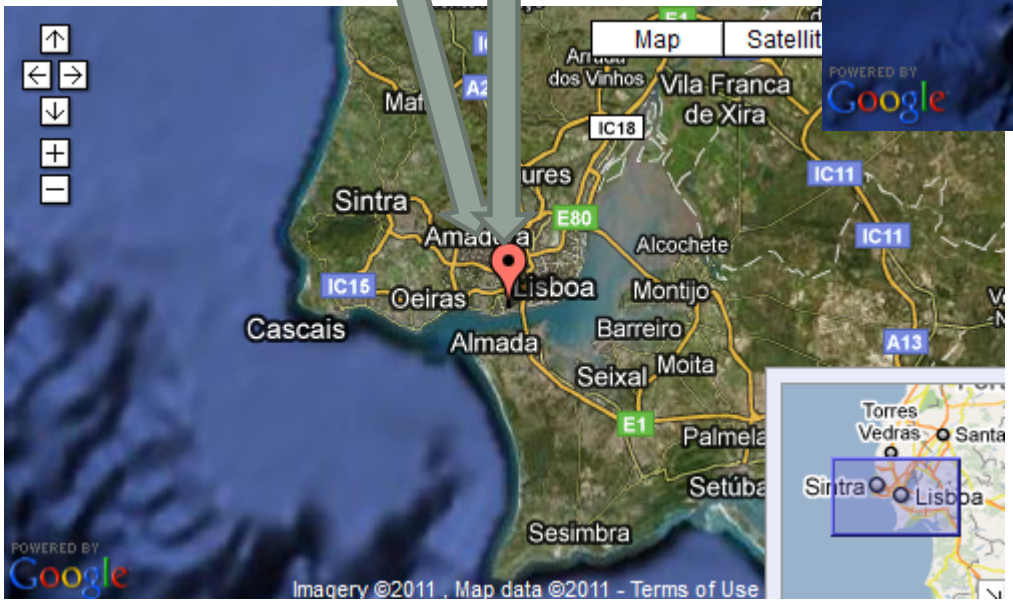
Some available tools – geographic

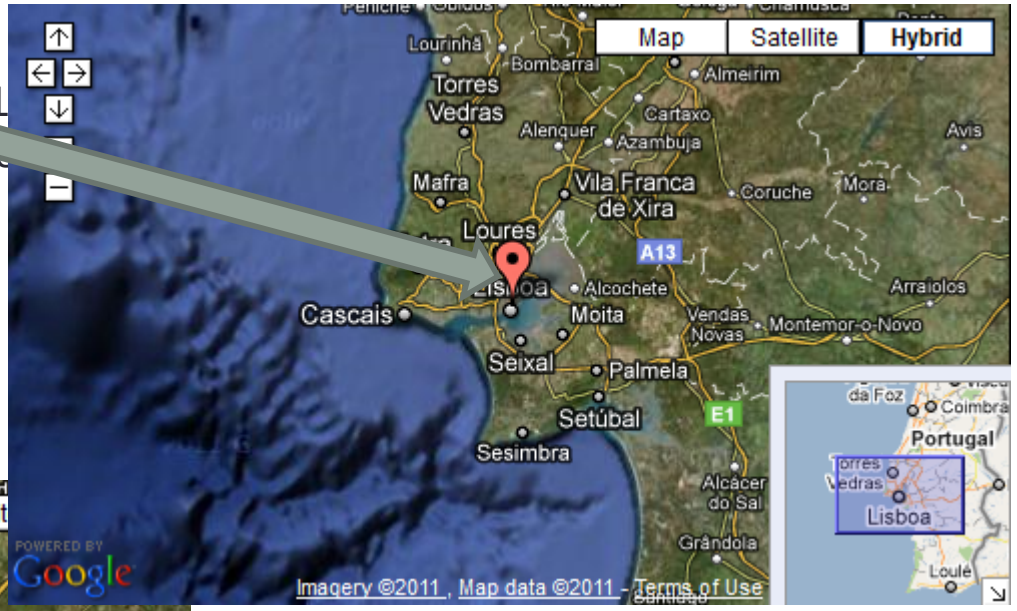
- Yahoo! Placemaker
  - Place name recognition and resolution
  - http://developer.yahoo.com/geo/placemaker/

  - EuropeanaConnect Geoparser
  - Place name recognition and resolution
  - Supports DC and ESE metadata
  - http://europeana-geo.isti.cnr.it/geoparser

The EuropeanaConnect Geoparser

identifier: BN-E. 4639 P.
title: **Belem Castle**, **Lisbon**
creator: Stanfield, William Clarkson, 1793-1
description: View da **Torre de Belém** rode
pescadores a puxar as redes
publisher: J.Murray
publisher: London
contributor: Finden, Edward, 1791-1857
date: 1832



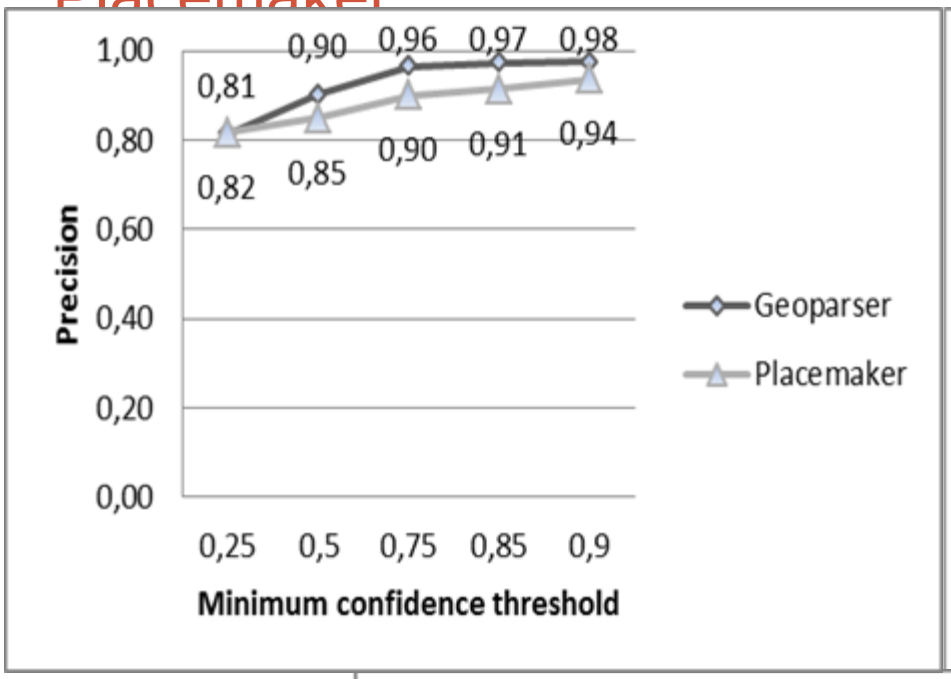


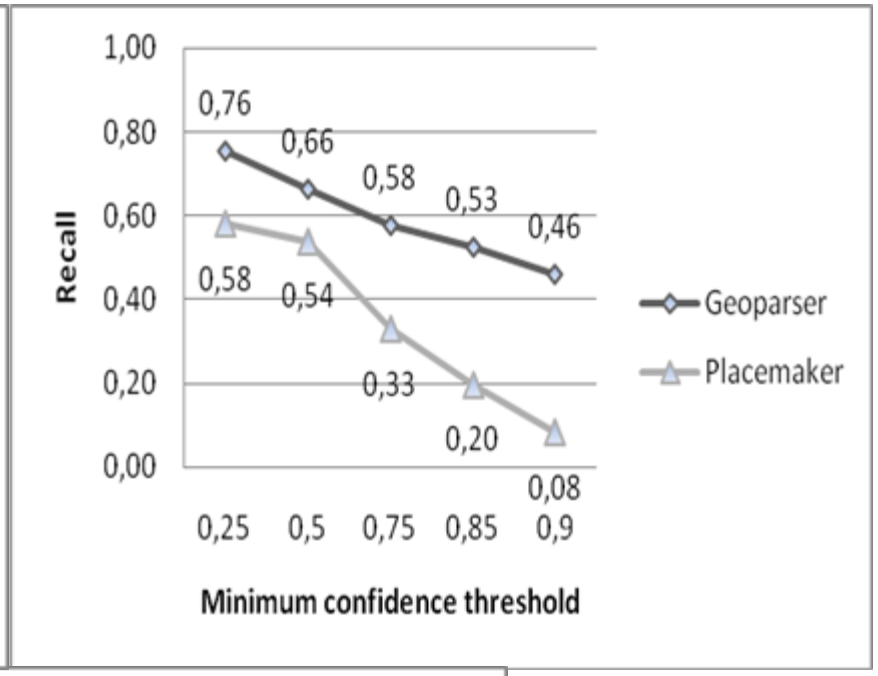
Comparison EuropenaConnect Geoparser - Yahoo! Placemaker

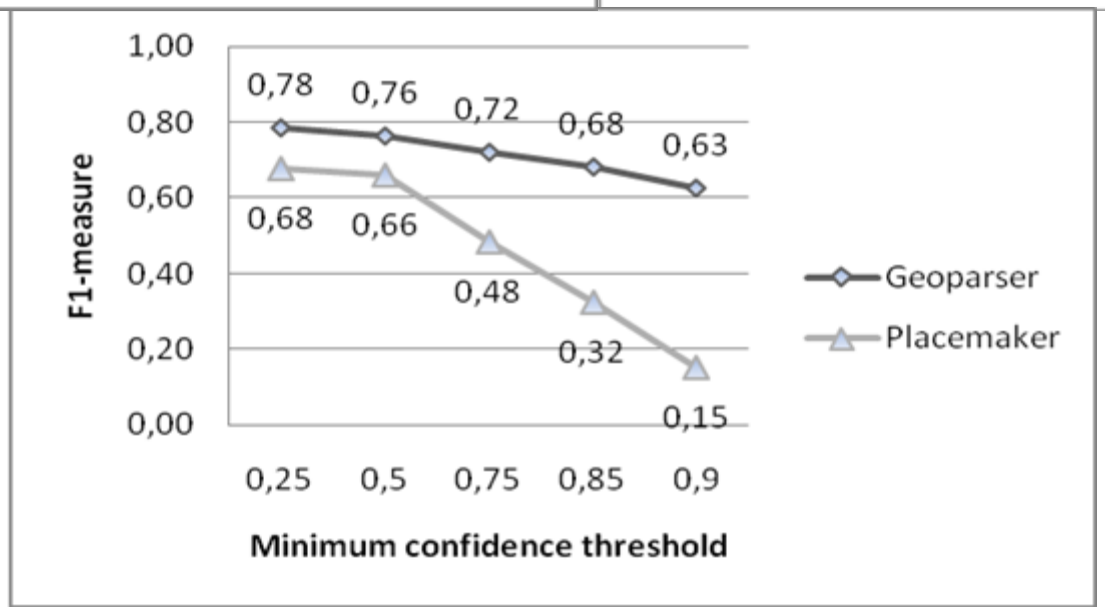
- Evaluation was performed on 752 ESE records
  - Records contained at least one place name
  - Maximum 20 records from a data provider/collection
  - All records were manually annotated (geoparsed)

- This collection was processed by both systems and the results were compared

Comparison EuropenaConnect Geoparser - Yahoo! Placemaker







Conclusions

- Information extraction tools are usually developed for well-structured text, and are language dependent.
- Information extraction involves a level of uncertainty:
  - Ambiguity is often present
  - Evidence may be insufficient
- Adopted tools usually provide a confidence value on the extracted information
  - Applications must choose the appropriate balance between precision and recall for their purposes

Thank you for your attention

Questions or comments?

Contact:

Nuno Freire – nfreire@gmail.com

Geoparser prototype (Europeana Semantic Elements)

http://europeana-geo.isti.cnr.it/geoparser