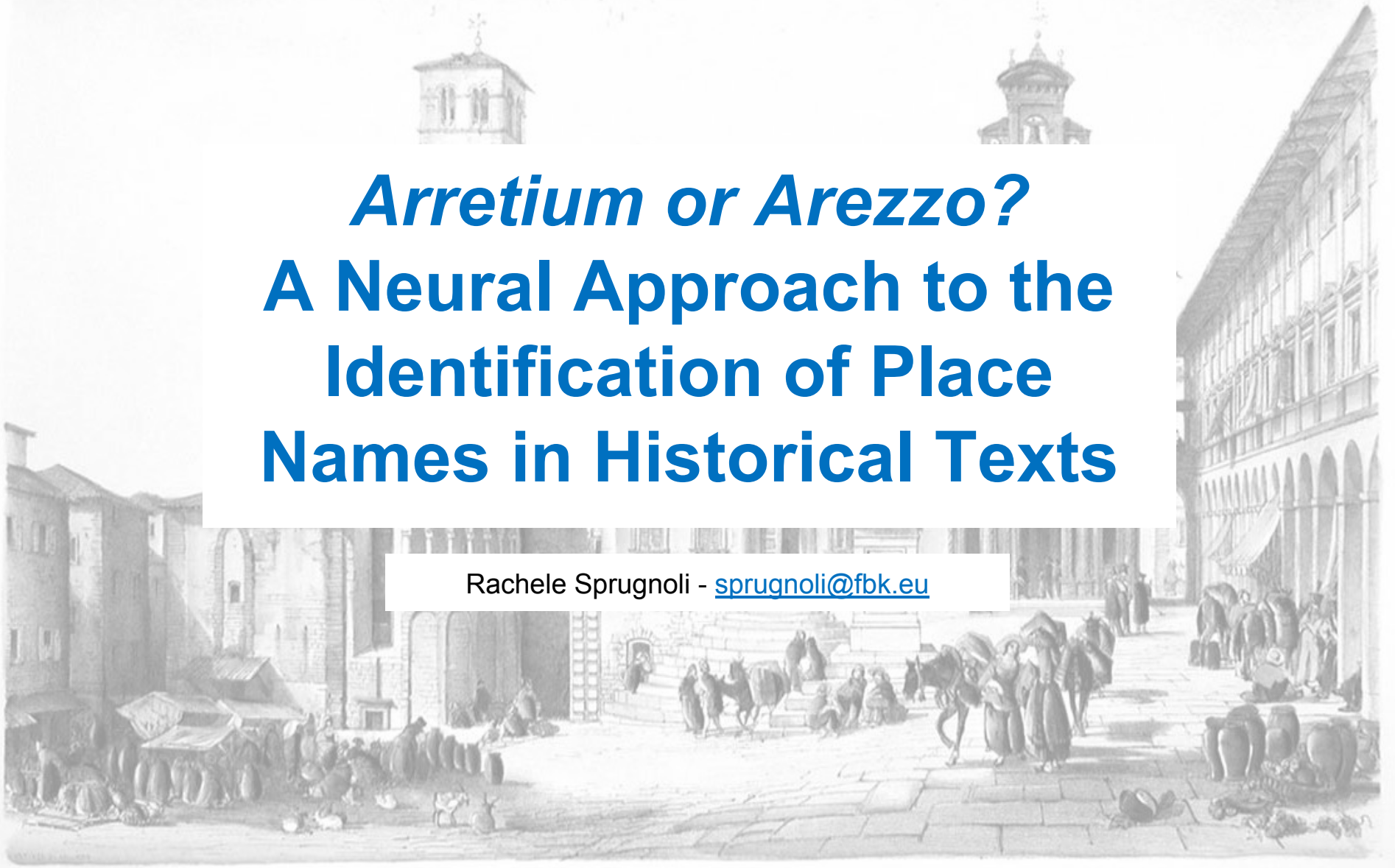
FONDAZIONE BRUNO KESSLER

DH DIGITAL HUMANITIES



Arretium or Arezzo? A Neural Approach to the Identification of Place Names in Historical Texts

Rachele Sprugnoli - sprugnoli@fbk.eu

WHAT

- Automatic extraction of place names in English historical travel writings

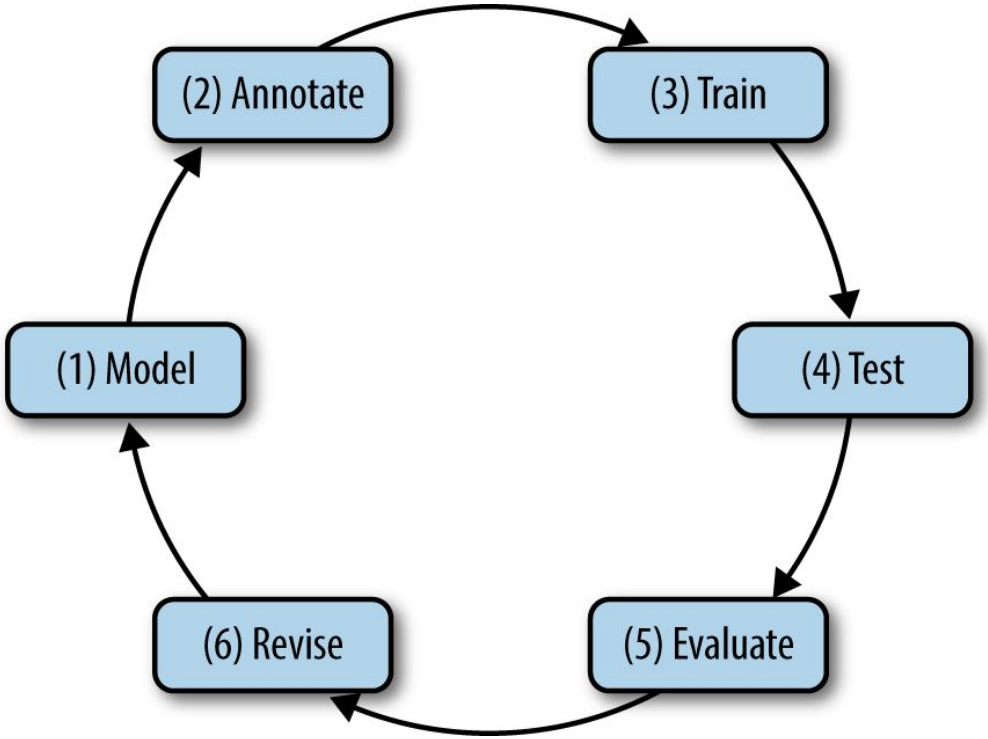
- Contributions:
  - release of a new annotated corpus
  - release of new historical word embeddings
  - release of the best model based on a neural architecture

WHY

- Research questions:
  - Are state-of-the-art NER systems good to identify place names in historical texts?
  - If not, can we develop a better model?
- Hot topic in the DH community:
  - spatial humanities framework

→ **Foster the collaboration between NLP and DH**

HOW



From model to annotated data to automatic system

(Pustejovsky & Stubbs, 2012)

Definition of Place Names

- We focus on three types of locations:
  1. **Geographical**: *Vesuvius, Mediterranean Sea, Campagna Romana, Mars*
  2. **Political**: *Venice, Tuscany, Regno delle due Sicilie, Vatican*
  3. **Functional**: *Hotel Riposo, Church of St. Severo, Forum Romanum, Via dell'Indipendenza*

CAMPAGNA ROMANA

Definition of Place Names

- We focus on three types of locations:
  1. **Geographical**: *Vesuvius, Mediterranean Sea, Campagna Romana, Mars*
  2. **Political**:*Brescia, Tuscany, Regno delle due Sicilie, Vatican*
  3. **Functional**: *Hotel Riposo, Church of St. Severo, Forum Romanum, Via dell'Indipendenza*

BRESCIA

Definition of Place Names

- We focus on three types of locations:
  1. **Geographical**: *Vesuvius, Mediterranean Sea, Campagna Romana, Mars*
  2. **Political**: *Venice, Tuscany, Regno delle due Sicilie, Vatican*
  3. **Functional**: *Hotel Riposo, Church of St. Severo, Forum Romanum, Via dell'Indipendenza*

FORUM ROMANUM

Manual Annotation

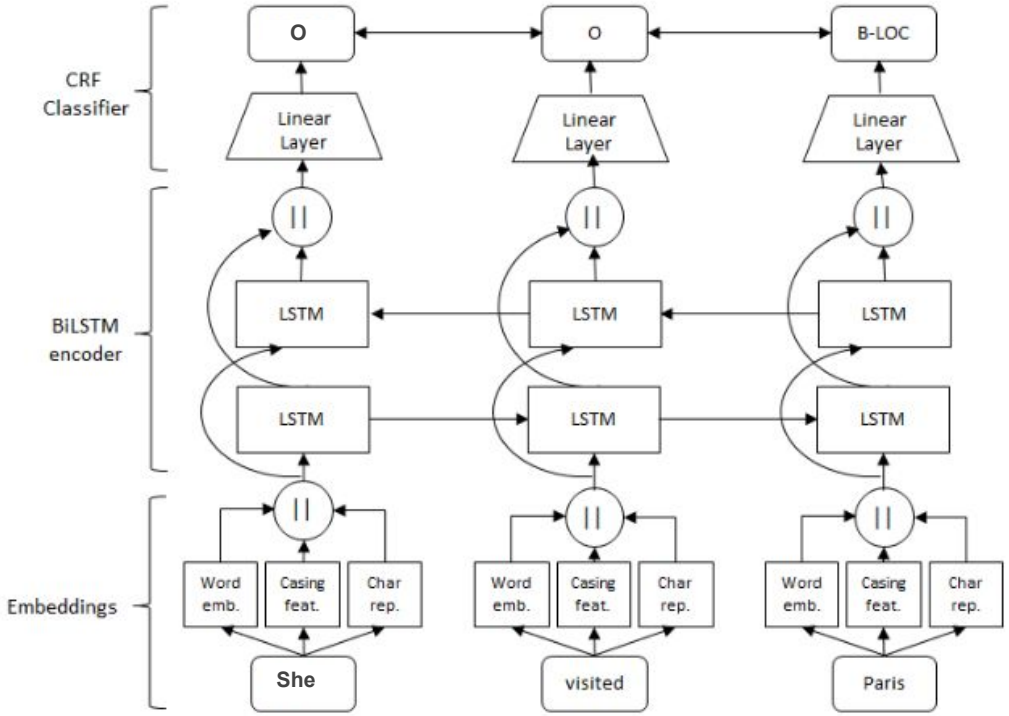
- Corpus manually annotated with place names:
  - English travel writings on Italy: 1860-1930
  - 38 texts from https://sites.google.com/view/travelwritingsonitaly
  - 100,000 tokens
  - 2,228 place names
  - IAA: 0.93 Cohen's kappa

Specific characteristics of place names

- Major issues when dealing with place names in travel writings:
  - Spelling variations: *Trapani* & *Drepanum* / *Venice* & *Venezia*
  - Presence of Latin graphemes: *Ætna*
  - Wrong spelling: *Cammaiore* instead of *Camaiore*
  - Long multi-token names: *House of the Tragic Poet*
  - Abbreviated forms: *Hotel B.*
  - High diversity: from *Val Buona* to *Capo S. Vito*
  - References to places outside Italy: *Savoie*

Experiments

- Corpus split: 80/10/10
- 3 steps:
  1) Testing of Stanford CoreNLP NER module
  2) Retraining of Stanford CoreNLP NER module
  3) Training and testing of a Bi-LSTM architecture



Reimers & Gurevych, 2017

Neural Approach: BiLSTM

- Best setup for the NER task suggested by Reimers & Gurevych:
  - classifier: CRF
  - character embeddings: CNN
  - word embeddings: GloVe Common Crawl 840B

Other pre-trained word embeddings:

- Levy and Goldberg
- fastText
- HistoGlove
- HistoFast
- HistoLevy

Historical Word Embeddings

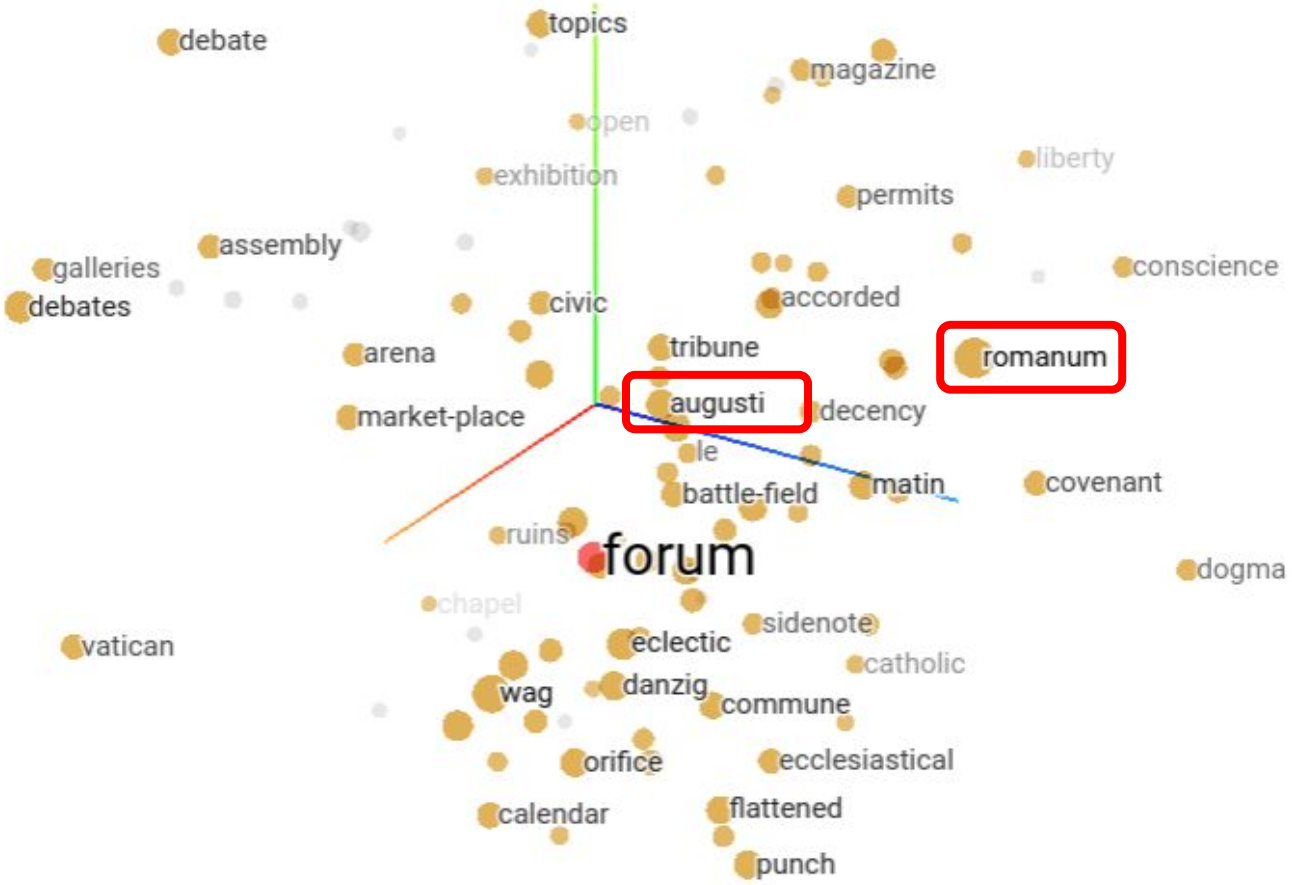
- From a subset of Corpus of Historical American English (COHA)

Glove

Historical Word Embeddings

- From a subset of Corpus of Historical American English (COHA)

HistoGlove



Results

	P	R	F1
Stanford NER	82.1	66.1	73.2
Retrained Stanford NER	78.9	79.2	79.1
Neural HistoLevy	85.3	83.3	84.3
Neural Levy	83.7	86.8	85.3
Neural HistoFast	83.9	87.4	86.0
Neural GloVe	83.7	87.9	86.0
Neural FastText	86.3	86.3	86.3
Neural HistoGlove	**86.4**	**88.5**	**87.4**

Results: details on test data

	P	R	F1
Stanford NER	82.1	66.1	73.2
Retrained Stanford NER	78.9	79.2	79.1
Neural HistoLevy	85.3	83.3	84.3
Neural Levy	83.7	86.8	85.3
Neural HistoFast	83.9	87.4	86.0
Neural GloVe	83.7	87.9	86.0
Neural FastText	86.3	86.3	86.3
Neural HistoGlove	**86.4**	**88.5**	**87.4**

	Stanford NER	Neural HistoGlove
	F1	F1
Little Pilgrimage, 1911	80.9	90.7
Naples Riviera, 1907	73.3	86.0
Rome, 1905	55.6	80.9

Conclusions & Future Works

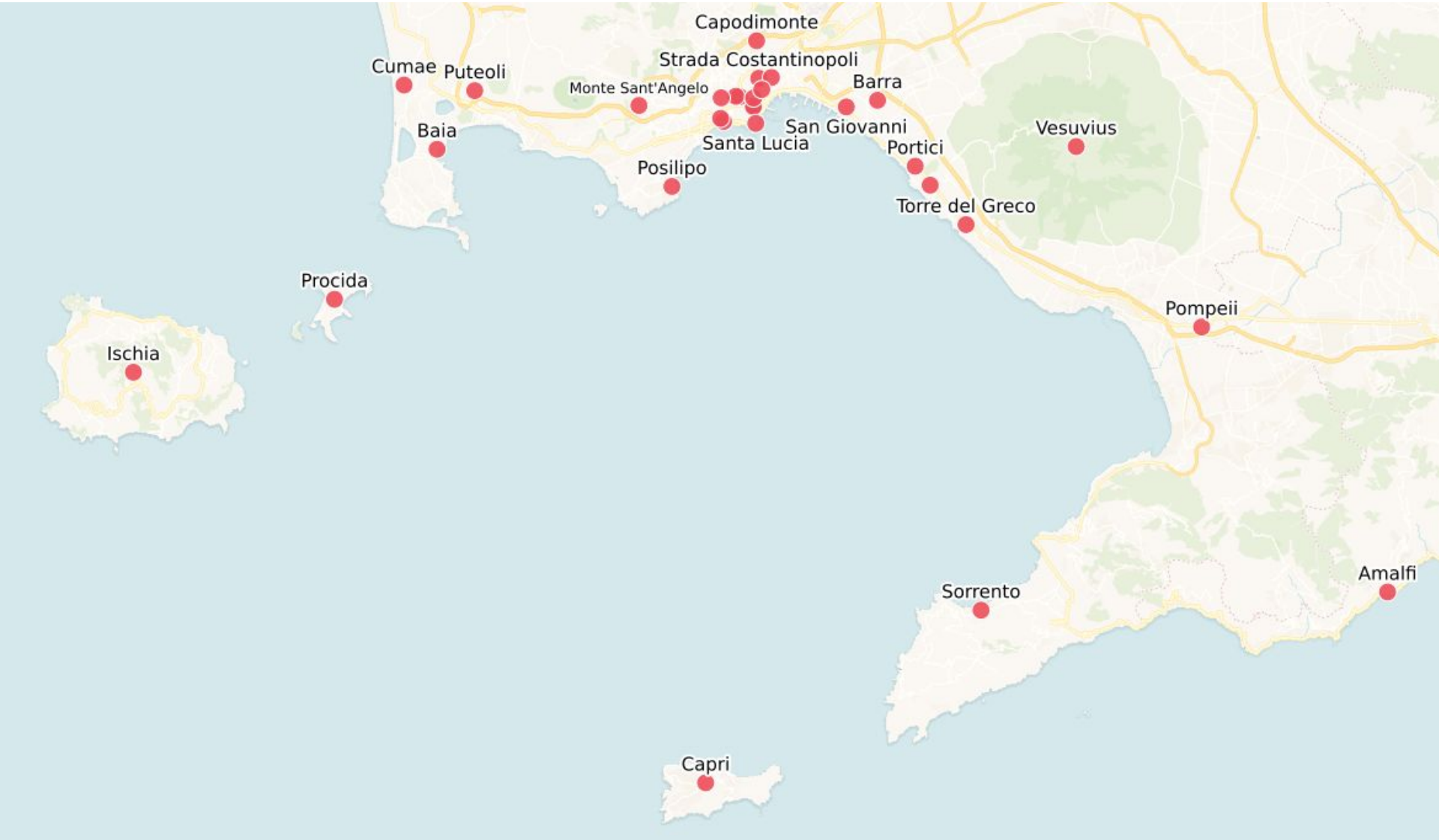
- Neural approach + in-domain annotated data + historical embeddings

- Available online: http://dh.fbk.eu/technologies/place-names-historical-travel-writings

- TO DO:
  - finer-grained classification
  - itinerary detection
  - geocoding

Example of Geocoding

Place names extracted from “Naples Riviera” test file



THANK YOU!

Email: sprugnoli@fbk.eu
Web Site: http://dh.fbk.eu
Twitter: @DH_FBK