NLP AND IR METHODS FOR HANDLING GEOSPATIAL INFORMATION IN TEXTUAL DOCUMENTS

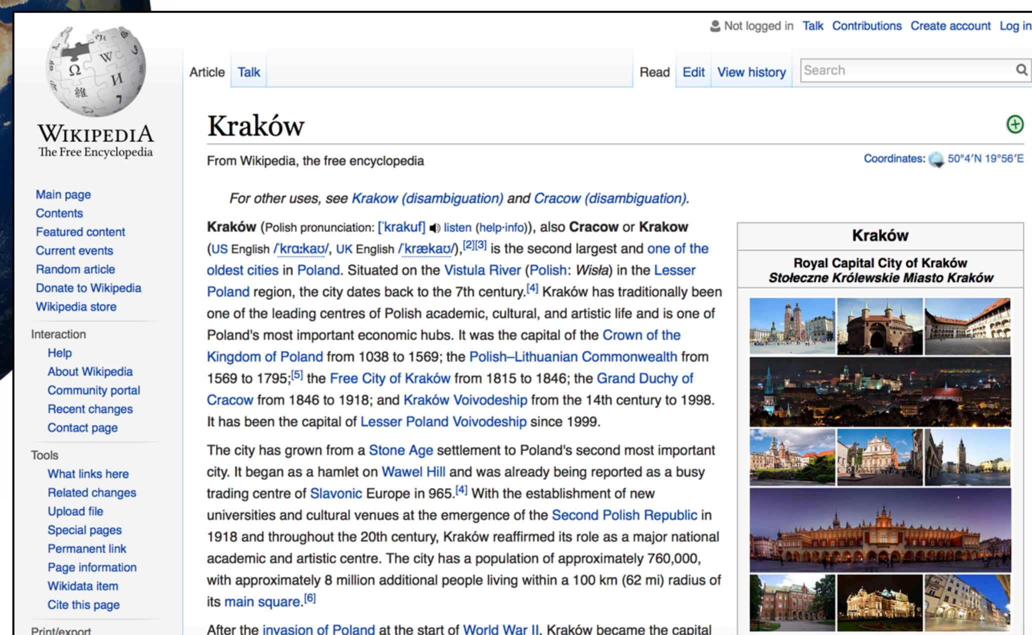
BRUNO MARTINS

JULY 11TH, 2016

DOCUMENT GEOCODING

LINKING DOCUMENTS TO GEOSPATIAL COORDINATES

?



Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export

Kraków

From Wikipedia, the free encyclopedia

Coordinates: 50°4′N 19°56′E

For other uses, see Krakow (disambiguation) and Cracow (disambiguation).

Kraków (Polish pronunciation: [ˈkrakuf] listen (help·info)), also **Cracow** or **Krakow** (US English /ˈkrɑːkaʊ/, UK English /ˈkrækaʊ/),[2][3] is the second largest and one of the oldest cities in Poland. Situated on the Vistula River (Polish: *Wisła*) in the Lesser Poland region, the city dates back to the 7th century.[4] Kraków has traditionally been one of the leading centres of Polish academic, cultural, and artistic life and is one of Poland's most important economic hubs. It was the capital of the Crown of the Kingdom of Poland from 1038 to 1569; the Polish–Lithuanian Commonwealth from 1569 to 1795;[5] the Free City of Kraków from 1815 to 1846; the Grand Duchy of Cracow from 1846 to 1918; and Kraków Voivodeship from the 14th century to 1998. It has been the capital of Lesser Poland Voivodeship since 1999.

The city has grown from a Stone Age settlement to Poland's second most important city. It began as a hamlet on Wawel Hill and was already being reported as a busy trading centre of Slavonic Europe in 965.[4] With the establishment of new universities and cultural venues at the emergence of the Second Polish Republic in 1918 and throughout the 20th century, Kraków reaffirmed its role as a major national academic and artistic centre. The city has a population of approximately 760,000, with approximately 8 million additional people living within a 100 km (62 mi) radius of its main square.[6]

After the invasion of Poland at the start of World War II, Kraków became the capital

Kraków

Royal Capital City of Kraków
Stołeczne Królewskie Miasto Kraków

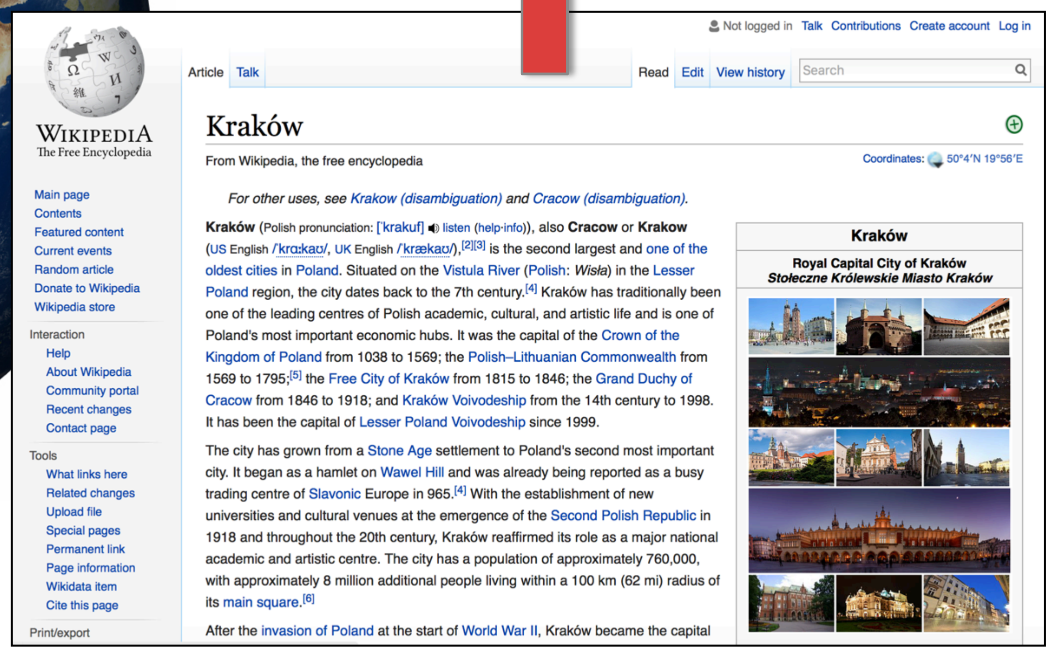
Wikipedia page for "Kraków"

DOCUMENT GEOCODING

LINKING DOCUMENTS TO GEOSPATIAL COORDINATES

Latitude : 50°03'41"N

Longitude : 19°56'18"E



Not logged in Talk Contributions Create account Log in

Article Talk Read Edit View history Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export

Kraków

From Wikipedia, the free encyclopedia

Coordinates: 50°4′N 19°56′E

For other uses, see Krakow (disambiguation) and Cracow (disambiguation).

Kraków (Polish pronunciation: [ˈkrakuf] listen (help·info)), also **Cracow** or **Krakow** (US English /ˈkrɑːkaʊ/, UK English /ˈkrækaʊ/),[2][3] is the second largest and one of the oldest cities in Poland. Situated on the Vistula River (Polish: *Wisła*) in the Lesser Poland region, the city dates back to the 7th century.[4] Kraków has traditionally been one of the leading centres of Polish academic, cultural, and artistic life and is one of Poland's most important economic hubs. It was the capital of the Crown of the Kingdom of Poland from 1038 to 1569; the Polish–Lithuanian Commonwealth from 1569 to 1795;[5] the Free City of Kraków from 1815 to 1846; the Grand Duchy of Cracow from 1846 to 1918; and Kraków Voivodeship from the 14th century to 1998. It has been the capital of Lesser Poland Voivodeship since 1999.

The city has grown from a Stone Age settlement to Poland's second most important city. It began as a hamlet on Wawel Hill and was already being reported as a busy trading centre of Slavonic Europe in 965.[4] With the establishment of new universities and cultural venues at the emergence of the Second Polish Republic in 1918 and throughout the 20th century, Kraków reaffirmed its role as a major national academic and artistic centre. The city has a population of approximately 760,000, with approximately 8 million additional people living within a 100 km (62 mi) radius of its main square.[6]

After the invasion of Poland at the start of World War II, Kraków became the capital

Kraków

Royal Capital City of Kraków
Stołeczne Królewskie Miasto Kraków

Wikipedia page for "Kraków"

TOPONYM RESOLUTION

LINKING INDIVIDUAL PLACE NAMES TO GEOSPATIAL COORDINATES

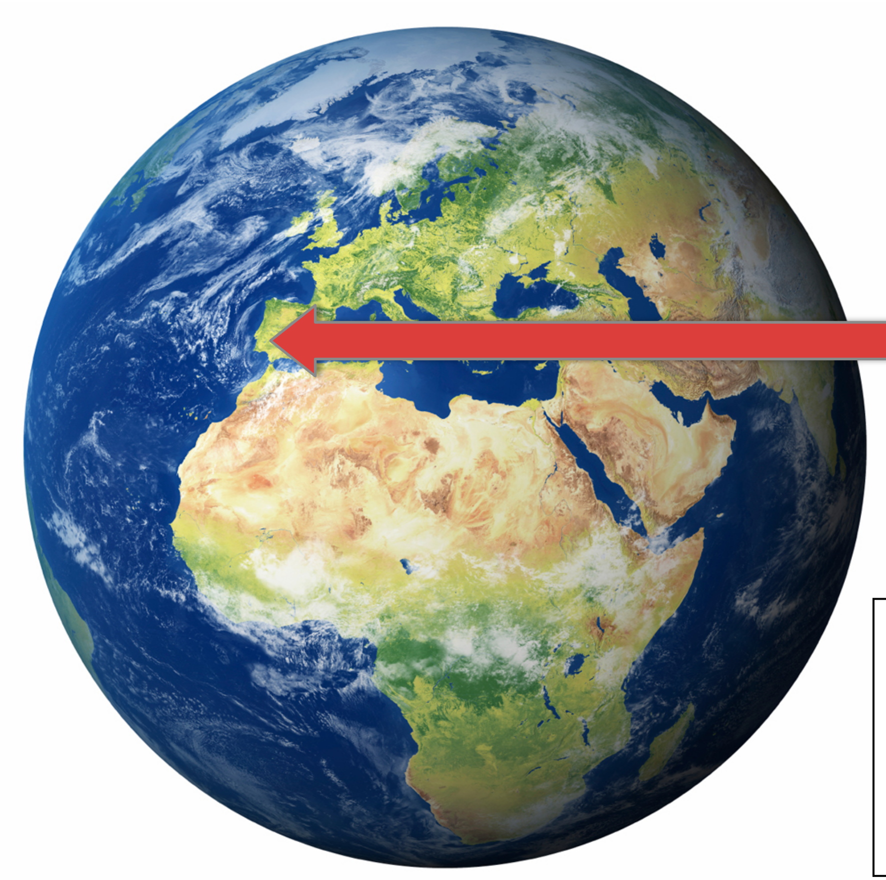
?

Kraków's historic centre, which includes the Old Town, Kazimierz and the Wawel Castle, was included as the first of its kind on the list of UNESCO World Heritage Sites in 1978.[73]

Wikipedia page for "Kraków"

TOPONYM RESOLUTION

LINKING INDIVIDUAL PLACENAMES TO GEOSPATIAL COORDINATES



Latitude : 50°3'15.98"N

Longitude : 19°56'11.69"E

Kraków's historic centre, which includes the Old Town, Kazimierz and the Wawel Castle, was included as the first of its kind on the list of UNESCO World Heritage Sites in 1978.[73]

Wikipedia page for "Kraków"

HANDLING GEOSPATIAL INFORMATION IN TEXT

- **Text and GIS Increasingly combined within DH research**
  - Cartographic visualization of information in document collections
  - Document retrieval according to geospatial constraints
  - Cross-links between resources
  - Spatial Humanities Project, Pelagios Project *(i.e., Pleiades+Peripleo+Recogito)*

- **Most previous work leverages *gazetteer matching*, together with *heuristics* for resolving ambiguous toponyms**
  - Place prominence, relations towards other places in same document

- **Challenges**
  - Gazetteer coverage *(e.g., vague regions, vernacular places, complete metadata)*
  - Toponym ambiguity *(i.e., geo/geo or geo/non-geo)*
  - Toponyms change over time, different spellings, different borders, …

- **State of the art methods from the NLP/IR communities still rarely considered in this practical application domain**

OVERVIEW

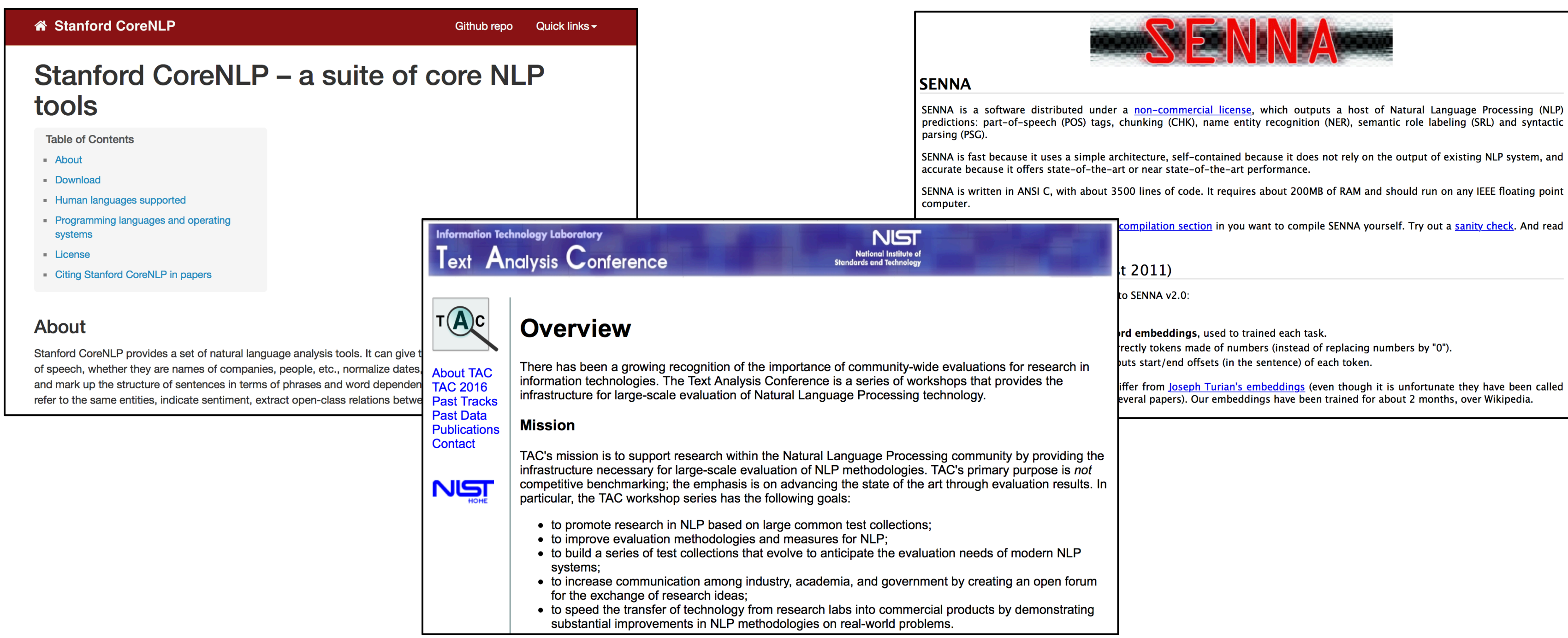
1. *Introduction and motivation*
2. **Modern NLP/IR methods**
    - Named entity recognition
    - Entity disambiguation
3. **Language modeling methods**
4. **Conclusions**

NAMED ENTITY RECOGNITION

- **Delimiting spans of text that correspond to entities**
- **Within the NLP community the task is modeled as a sequence classification/tagging problem**
- **Models are learned from labeled sequences, and they can then assign probabilities to tagging decisions (and, consequently, also to sequences of tags)**
  - Hidden Markov Models
  - Conditional Random Fields
  - Deep Neural Networks (e.g., CNNs, RNNs, …)
- **Current trends:** *avoid hand-engineered features, word embeddings, generalize across languages and domains*

NAMED ENTITY RECOGNITION RESOURCES

- Stanford Core NLP and Stanford NER
- SENNA and systems inspired on SENNA
- Competition at the Text Analysis Conference

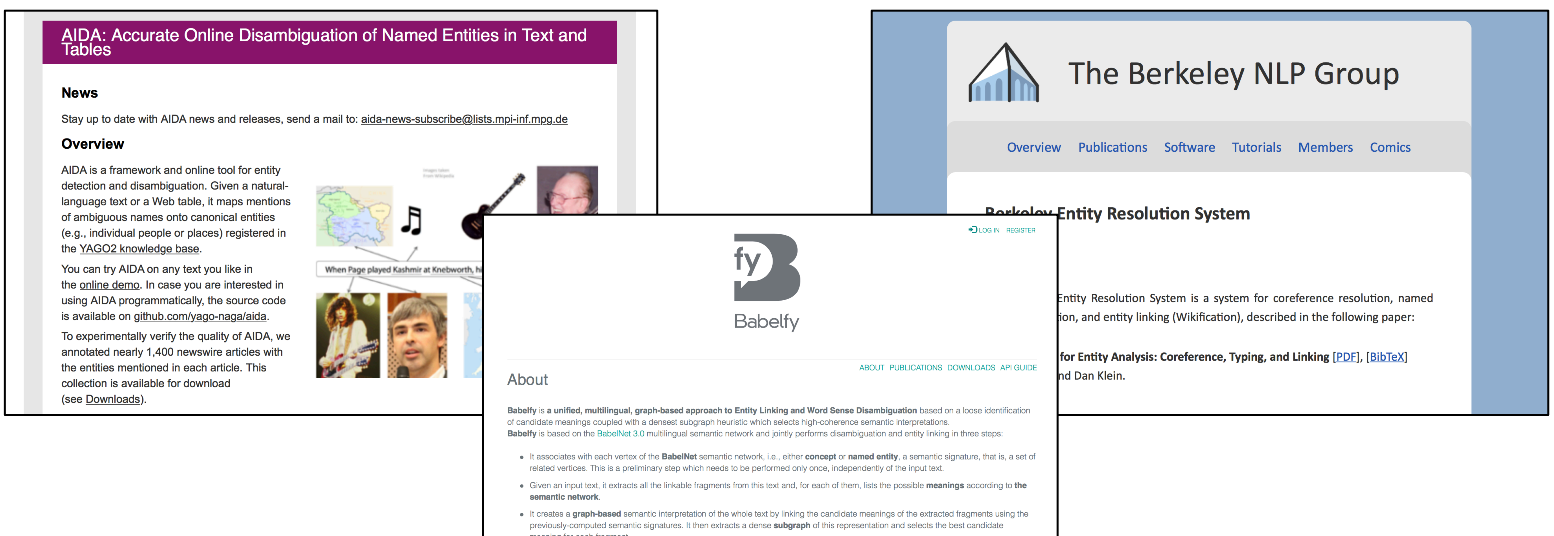


NAMED ENTITY DISAMBIGUATION

- **Link entities to a reference database (DB)**
- **Task is typically modeled as a candidate ranking problem, often also leveraging Wikipedia as the reference DB**
  - **Retrieve candidate disambiguations from a database**
    - Matching strings by similarity against Wikipedia concept names
  - **Rank according to likelihood of correct disambiguation**
    - ***Prior probability*** P(candidate|mention) from resources like Wikipedia
    - ***Context similarity*** between candidate and mention/document
    - ***Coherence*** between candidates (within same document)
    - Learn from examples to combine evidence and assign probability to candidates
- **Current trends:** *global disambiguation, concept/entity embeddings*
- **Several studies proposed heuristics specific for toponyms**
  - Work my Mike Lieberman, Jochen Leidner, …
  - Population, geospatial distance, …

NAMED ENTITY DISAMBIGUATION RESOURCES

- **AIDA/YAGO**
- **Babelfy** ***(entity linking and word sense disambiguation)***
- **Berkeley Entity Resolution** ***(handles co-references)***
- **Competition at the Text Analysis Conference**



MODERN NLP/IR METHODS

- **Discussed methods handling named entities in general**
  - Provide very good performance on toponyms
    - Named entity recognition : accuracy around 90%
    - Entity linking : accuracy around 80%
  - Portable across tasks, languages, domains, …
  - Methods actively developed in the NLP community, which now embraces open research and reproducibility of results
  - Robust software (although difficult to use by non experts)

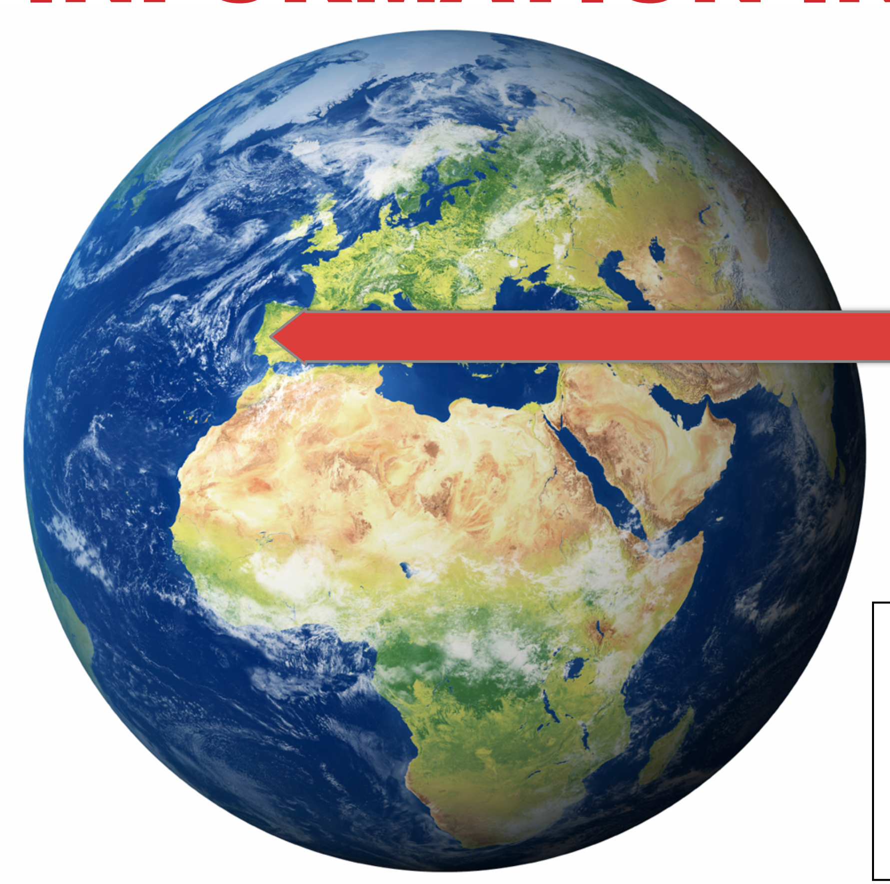
- **Even if recognition leverages patterns in annotated data, disambiguation still depends on reference DB**

- **Some studies have specifically focused on handling toponyms and geospatial information…**

OVERVIEW

1. *Introduction and motivation*
2. *Modern NLP/IR methods*
    - *Named entity recognition*
    - *Entity disambiguation*
3. **Language modeling methods**
4. **Conclusions**

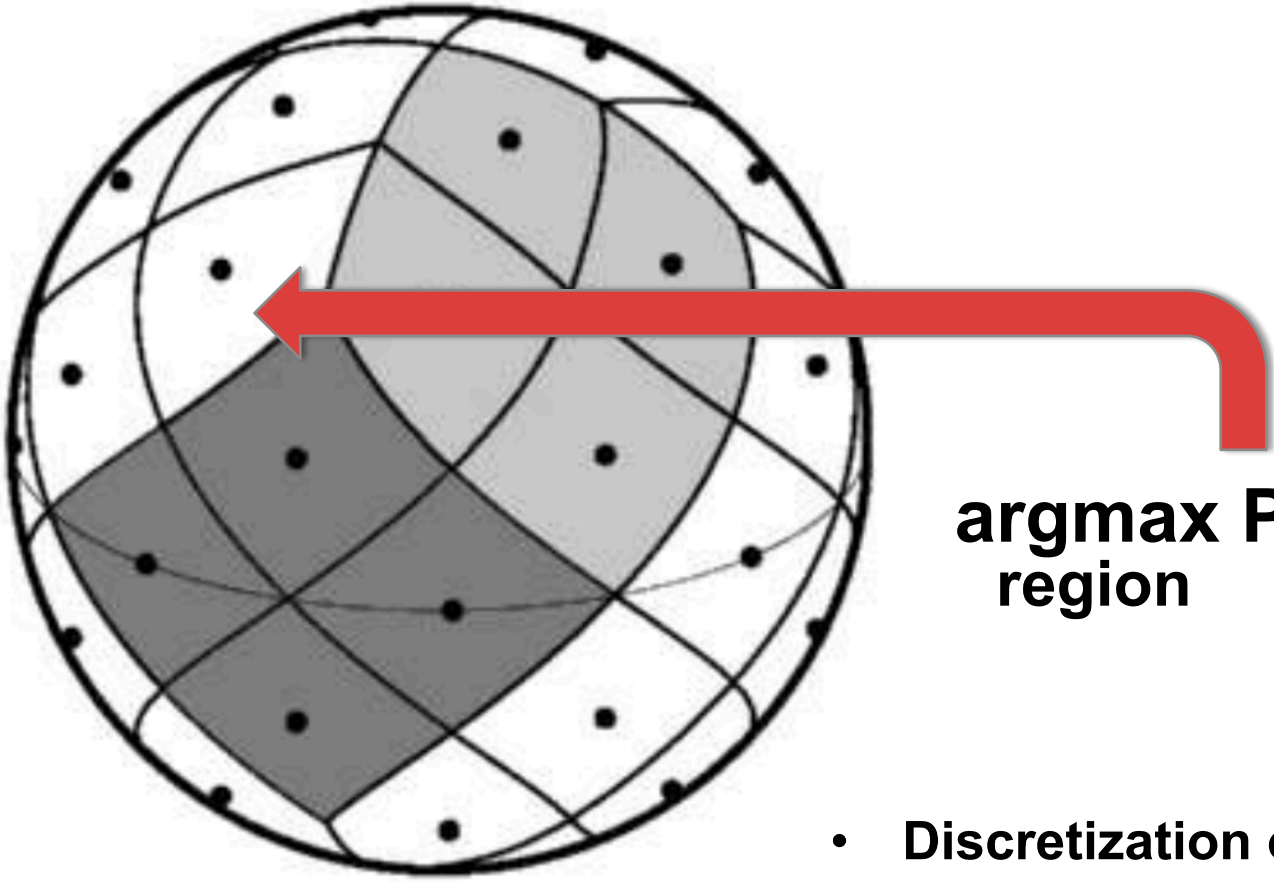
HANDLING GEOSPATIAL INFORMATION IN TEXTS



Kraków's historic centre, which includes the Old Town, Kazimierz and the Wawel Castle, was included as the first of its kind on the list of UNESCO World Heritage Sites in 1978.[73]

Wikipedia page for "Kraków"

AN APPROACH BASED ON LANGUAGE MODELING



$$\underset{region}{\operatorname{argmax}}\, P(region|text)$$

- Discretization of space
- Large datasets (e.g., Wikipedia)
- Standard language models

RELATED WORK
DOCUMENT GEOCODING

- **Several recent proposals based on language models (e.g., work by Baldridge et al.)**
  - **Discretize the surface of the Earth**
    - Regular grids versus hierarchical triangular meshes
  - **Train language models for each region of the discretization, with basis on available data (requires large datasets)**
    - Naïve Bayes models
    - Smoothed n-gram models
    - Discriminative classification methods
    - Neural language models (CNNs, RNNs, …)
  - Assign region(s) most likely to generate test document
  - Many other variations (*e.g., smoothing, term selection, …*)

RELATED WORK
DOCUMENT GEOCODING

We have exhaustively surveyed previous work in the area…

Significant progress over the years…

Author	Dataset	Method	Median dist.
Baldridge et al. 2011	Wiki EN	Unigram LM + KL div.	11,8 km
Baldridge et al. 2011	Twitter S	Unigram LM + KL div.	479,0 km
Baldridge et al. 2012	Wiki EN	K-d-tree + regular + NN	13,4 km
Baldridge et al. 2012	Twitter L	K-d-tree + NN	463,0 km
Laere et al. 2014	Wiki UK	K-medoids + feat. select.	4,2 km
Han et al. 2014	Twitter XL	IGR feature selection	640,0 km
Baldridge et al. 2014	Wiki EN	Logistic regression	15,3 km
Baldridge et al. 2014	Twitter XL	Logistic regression	490,0 km

Automated Geocoding of Textual Documents

A Survey of Current Approaches

Fernando Melo and Bruno Martins
{fernando.melo,bruno.g.martins}@ist.utl.pt

Instituto Superior Técnico and INESC-ID
Universidade de Lisboa, Portugal

Abstract
Geographical Information Retrieval (GIR) has captured the attention of many different researchers that work in fields related to language processing and to the retrieval and mining of relevant information from large document collections. With the rise of unstructured information being published on-line, we have also witnessed an increased interest in applying computational methods to extract geographic information from heterogeneous and unstructured data, including textual documents. This survey article specifically focuses on previous research addressing text-based document geocoding, i.e., the task of predicting the geospatial coordinates of latitude and longitude, that best correspond to a given document, with basis on its textual contents. We describe (i) early document geocoding systems that use heuristics over place names mentioned in the text (e.g., names of cities and states), (ii) probabilistic language
… world (usually considering a
…h are then used to predict per-
…f different heuristics, such as
…ge models built from different
…models.

…sification; Processing Geospatial

…aid to be related to some form of
…l information being published on-
…to extract geographic information

…of many different researchers that
…ilities of traditional geographical
…tions to formal geospatial coordi-
…essions (Purves and Jones, 2011;
…rmation extraction communities,
…ntities, specifically attempting to
…dependent. This includes studies
…l., 2010; Santos et al., 2015; Spe-
…c modeling (Speriosu et al., 2010;
…beyond named places (Liu et al.,
…l relations between places (Khan
… of spatial semantics from natural
…3).

RELATED WORK
TOPONYM RESOLUTION

- **Similar to document geocoding, considering text span around place reference**
    - *(often in combination with remaining text contained in the document, as back-off model)*

- **Avoid the use of gazetteers, instead relying on language models to better generalize**
    - Can handle vague geographic references (e.g., *downtown Kraków*)
    - Can handle relative references to places (e.g., *close to Kraków*)
    - Can assign text to multiple regions (e.g., raster representations)
    - ***Downside:*** Requires extensive amounts of training data

OVERVIEW

1. *Introduction and motivation*
2. *Modern NLP/IR methods*
    - *Named entity recognition*
    - *Entity disambiguation*
3. *Language modeling methods*
4. **Conclusions**

CONCLUSIONS

- **Reviewed related work on the NLP/IR communities**

- **Described simple procedure, based on language models, for assigning text to geospatial locations**
  - State-of-the-art results for document geocoding
  - Promising results in toponym resolution

- **Can leverage existing resources** (Wikipedia text)
- **Language and domain independent**
- **Easy to implement** (*out-of-the-box learning algorithms*)
- **Efficient and easy to parallelize**

- **Also easy to extend…**

MANY IDEAS FOR FUTURE WORK

- **Other statistical models and machine learning methods**
  - Novel neural network architectures
  - Structured sparsity (sentences, word clusters, …)
- **Experiments with other reference datasets**
  - Many previous studies have leveraged Wikipedia
  - Other datasets: Perseus Civil War collection, SpatialML
  - **The DH community can help significantly here**
- **Explore cross-language/domain correlations**
  - Much more data is available for English newswire text
- **Extensions and applications in other related tasks**
  - Assignment to geospatial regions instead of coordinates
  - Resolving trajectories described within documents
  - Extracting place characteristics and relations between entities and places

QUESTIONS?

BRUNO MARTINS

JULY 11TH, 2016

THANKS TO MY STUDENTS...

...

(they actually did most of the work!)