Text Meets Space: Geographic Content Extraction, Resolution and Information Retrieval

Jochen L. Leidner, Bruno Martins,
Katherine McDonough and Ross S. Purves

Tutorial held at ECIR 2020

Lisbon, Portugal, 14 April 2020
(online)

Tutorial Outline

- I. Geography and text
- **II. Toponym recognition and resolution** (J.L. Leidner)
- III. Geographic relevance and ranking
- IV. Applications
- V. Future challenges

Part II. Outline

- Introduction & Motivation
- Concepts & Terminology
- Place, Naming, Reference and Change
- (Address) Geocoding
- Gazetteers
- Named Entity Tagging
- Example Named Entity Tagger
- Toponym Resolution Heuristics
- Machine Learning-Based Methods
- Geographic Expression Parsing
- Example Toponym Resolver
- Annotation & Tools
- Text & Meta-Data
- Applications
- Where to Go from Here?

Introduction & Motivation

- Everything that happens, happens *somewhere*: time and space *situate* events.
- Geographic space provides a fundamental set of dimensions to order/structure information, as it structures the universe.
- Geographic meta-data (together with temporal meta-data) can usefully supplement content enrichment in order to improve information retrieval.

Place, Naming, Reference and Change

- Humans are place bound creatures, most of them have a place they call *home*.
- When humans migrate, they often even take their home place name with then: York - New York; Brunswick - New Brunswick
- Places: locations that have salience, therefore deserved to be named
- **Toponym**: name of a place (by linguistic convention, F. de Saussure (1916): arbitrariness of the sign)
- The name of a place signifies cultural connection:
  Londonjon/Lunden, Londinum, London;
  Saint Petersburg – Petrograd – Leningrad
- Geo-reference: objective (often numeric) way to refer to a location (e.g. grid reference)

Identifying Locations: Georeferencing Systems

Georeferencing System	Example
1 Placename (toponym)	**Edinburgh**
2 Postal address	Brandenburger Tor, Pariser Platz, Berlin
3 Postal code	CB2 1RD
4 Phone calling area	+1 (212) ...
5 Latitude/longitude	52.516272, 13.377722 (52°30′58.58″ N, 13°22′39.80″ E)
6 UK National Grid	NT 252 734

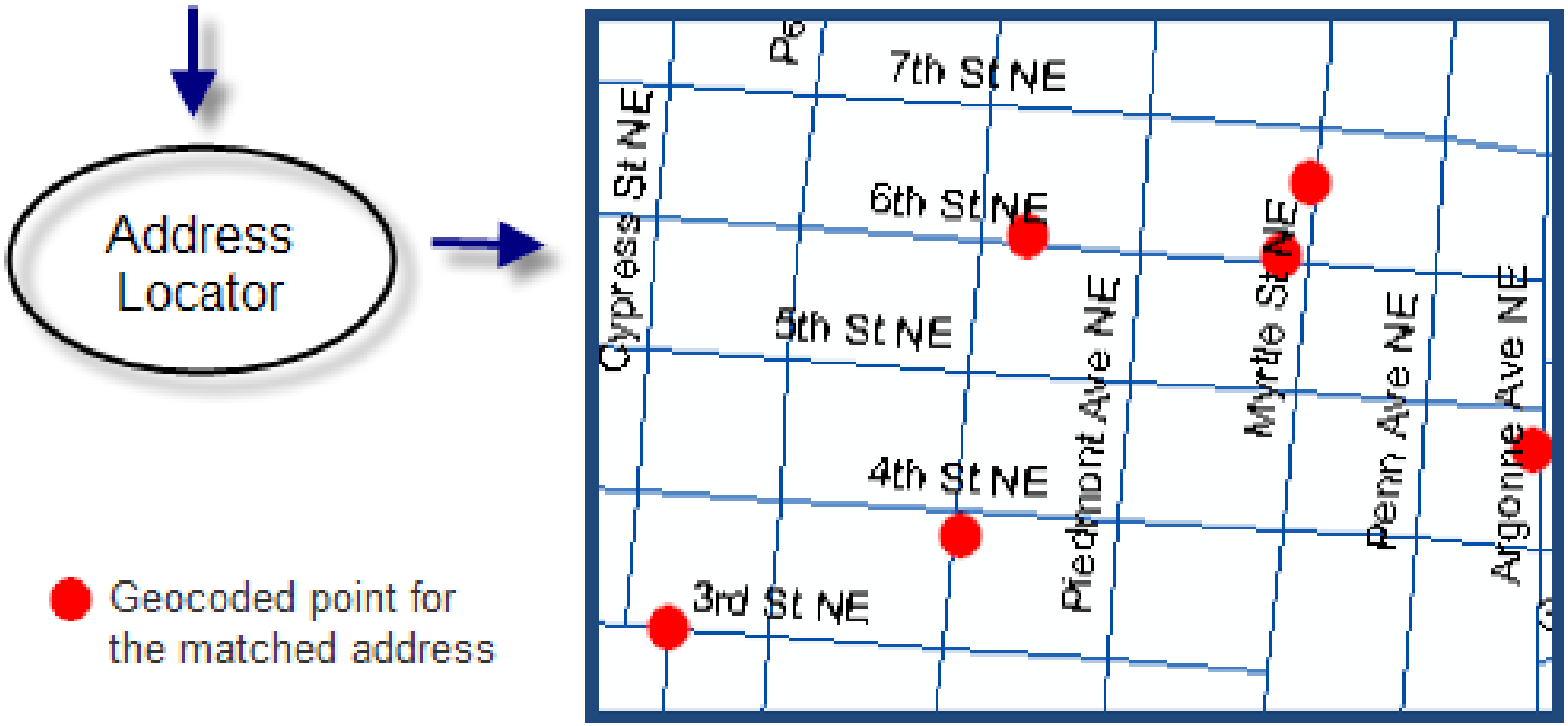
(See also Longley et al. (2005), Chapter 5 (Georeferencing), p. 107ff.) Translating Georeferences

Geocoding:	2	→	5
Toponym resolution:	1	→	5
Reverse geocoding:	5	→	1

ESRI ArcGIS

customers

	NAME	ADDRESS	CITY	S
▶	Ace Market	1171 PIEDMONT AVE NE	ATLANTA	G
	Andrew's Gasoline	1670 W PEACHTREE ST NE	ATLANTA	
	AP Supermarket	4505 BEVERLY RD NE		G
	Atlanta Market	241 16TH ST NW	ATLANTA	G



(Source: ESRI Inc.)

Gazetteers

- A **gazetteer** is a lexical resource (database) comprising (Hill, 2000): a set of
  - toponyms as keys
  - a **feature type**
  - a **spatial footprint**
- Example: "When we landed, we found our chaise ready, and passed through **Kinghorn, Kirkaldy, and Cowpar**, places not unlike the small or straggling **market-towns** in those parts of **England** where commerce and manufactures have not yet produced opulence."

  – Johnson, Samuel, *A journey to the western islands of Scotland*, Oxford Text Archive, ⟨`http://hdl.handle.net/20.500.12024/0076`⟩, page 4

  ⟨Kinghorn, town, NT271869⟩

Some Online Gazetteer Resources

Gazetteer Name	World Wide Web Location
Alexandria Gazetteer	http://www.alexandria.ucsb.edu/gazetteer
US CIA World Fact Book	https://www.cia.gov/cia/publications/factbook/index.html
Getty Thesaurus of Geographic Names	http://shiva.pub.getty.edu/tgn_browser/
US NGA GEOnet Names	http://164.214.2.59/gns/html/index.html
Ordnance Survey (OS) 1:50,000 Scale Gazetteer	http://www.ordnancesurvey.co.uk/oswebsite/products/
Seamless Administrative Boundaries of Europe (SABE)	http://www.eurogeographics.org/eng/03_projects_sabe.asp
United Nations (UNECE) UN-LOCODE	http://www.unece.org/cefact/
US Census Gazetteer	http://www.census.gov/cgi-bin/gazetteer/
US Geological Survey Geographic Names	http://www-nmd.usgs.gov/www/gnis/

Named Entity Tagger

- Also: name tagger, ner, ne tagger
- A piece of software for Named Entity Recognition and Classification
- Identify all text spans that mention proper nouns
- Classify the type of the entity named e.g. location (LOC), person (PER), organization (ORG), time expression (TIM), ...
- How does it work?
  - Human-written rules: word is capitalized, word to the right is 'major' $\Rightarrow$ LOC
  - Human-collected lexicons: e.g. Gittings (2012)
  - Statistics: $P(t = LOC|w = Scotland)$ estimated from hand-labeled training data-set
- Examples: TreeTagger, TnT, C&C, SpaCy, Stanford CoreNLP NE Tagger, Stanza, Refinitiv OpenCalais, GATE ANNIE, GermaNER.

Named Entity Tagging: Our Text (Johnson, *ibd.*, p.3)

As we crossed the #Frith$ of #Forth,$ our curiosity was attracted by #Inch #Keith,$ a small island, which neither of my companions had ever visited, though, lying within their view, it had all their lives solicited their notice. Here, by climbing with some difficulty over shattered crags, we made the first experiment of unfrequented coasts. Inch Keith is nothing more than a rock covered with a thin ((layer))

<P 3>

Named Entity Tagging: Our Text (Johnson, *ibd.*, p.25

"In old Aberdeen$_{\text{LOC}}$ stands the King's College$_{\text{ORG}}$, of which the first president was Hector Boece$_{\text{PER}}$, or Boethius$_{\text{PER}}$, who may be justly reverenced one of the revivers of elegant learning. When he studied at Paris$_{\text{ORG}}$, he was acquainted with Erasmus$_{\text{PER}}$, who afterwards gave him a public testimony of his esteem, by inscribing to him a catalogue of his works."

– Johnson, Samuel, *ibd.*, page 25

Toponym Ambiguity

- Ambiguity is a challenge:

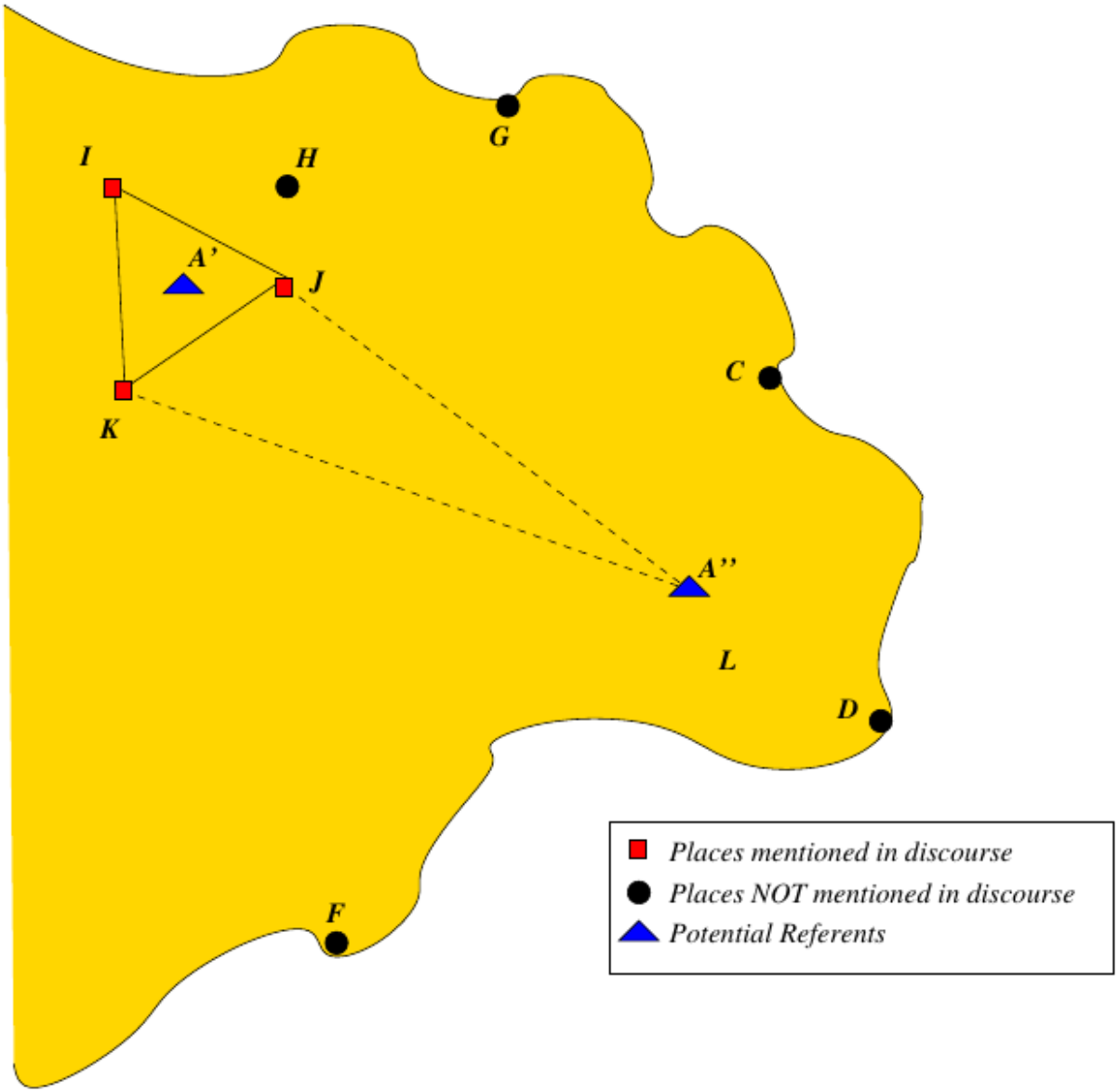
geo/geo ambiguity	Cambridge (MA, USA)	⟷	Cambridge (England, GB)
	Paris (TX, USA)	⟷	Paris (France)
geo/non-geo ambiguity	Of (Turkey)	⟷	Of (preposition)
	March (England, GB)	⟷	March (month)

- There are 1,625 "Santa Ana"s on earth.
- Humans typically have a notion of the most salient one, to the computer they look all the same.
- While I did a Ph.D. on toponym ambiguity, I heard about people flying to all the wrong places.

Toponym Resolution Heuristics (Leidner, 2007)

No.	Description
/H1/	'Contained-in' qualifier following
/H2/	Superordinate mention
/H3/	Largest population
/H4/	One referent per discourse
/H5/	Geometric minimality
/H6/	Singleton capitals
/H7/	Ignore small places
/H8/	Focus on geographic area
/H9/	Distance to unambiguous textual neighbours
/H10/	Discard off-threshold
/H11/	Frequency weighting
/H12/	Prefer higher-level referents
/H13/	Feature type disambiguators
/H14/	Textual-spatial correlation
/H15/	Default referent

Spatial Minimality (Leidner, 2007)



Spatial Minimality: Example (Leidner, 2007)

{ *Berlin*; Potsdam } ↦ Berlin, FRG (Germany)

{ Fairburn; *Berlin* } ↦ Berlin, WI, USA

{ West *Berlin*; Bishops; Dicktown } ↦ Berlin, NJ, USA

{ Kensington; *Berlin*; New Britain } ↦ Berlin, CT, USA

{ Copperville; *Berlin*; Gorham } ↦ Berlin, NH, USA

{ Moultrie; *Berlin* } ↦ Berlin, GA, USA

{ *Berlin*; Prouty } ↦ Berlin, IL, USA

{ *Berlin*; Berlin Center; Cherryplain } ↦ Berlin, NY, USA

{ Medberry; *Berlin* } ↦ Berlin, ND, USA

Toponym Resolution Using Machine Learning

- Supervised classifier for feature types (Garbin and Mani, 2005)
- Partial solutions: state/country classifiers (Smith and Mann, 2003)
- Co-occurrence based methods (Overell, 2009; DeLozier et al, 2015)
- **Topocluster**: Local Getis-Ord statistic (DeLozier et al., 2015)
- **CamCoder**: Convolutional Neural Network (Gritta, Pilehvar and Collier, 2018)

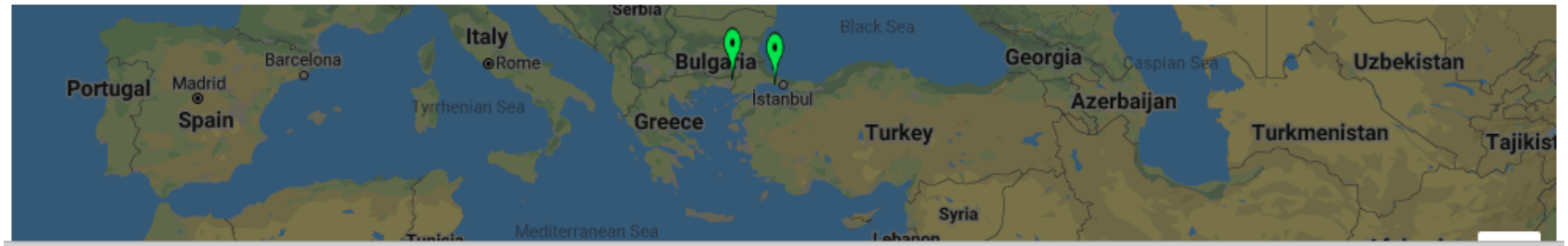
Geographic Expression Parsing

(1) *30 minutes north of Paris by car*
(2) *Clapham, a district south-west London lying mostly within the London Borough of Lambeth*
(3) *near Bruntsfield Links*
(4) *approximately seven kilometers from the German border*
(5) *halfway between Glasgow and Edinburgh*

(Source: Leidner, in print)

- Compositional syntactic analysis wanted
- Example: Bilhaut et al. (2003): Definite Clause Grammars (Prolog DCGs)

The Edinburgh Geoparser (Grover et al., 2010)



happens that many women die single at an advanced age, having never been able to fulfil the conditions required.

CXVII I. To these nations, which I have described, assembled in council, the Scythian ambassadors were admitted; they informed the princes, that the Persian, baring reduced under bis authority all the nations of the ad- joining continent, had thrown a bridge over the neck of the Bosphorus, in order to pass into theirs: that he had already subdued Thrace, and constructed a bridge over the Ister, am- bitiously hoping to reduce them also. "Will it be just," they continued, "for you to remain inactive spectators of our ruin? Rather, having the same sentiments, let us advance together against this invader: unless you do this, we shall be reduced to the last extremities, and be compelled either to forsake our country, or to submit to the terms he may impose. If you withhold your assistance, what may we not dread? Neither will you have reason to expect a different or a better fate: for are not you the object of the Persian's ambition as well as our- selves? or do you suppose that, having van- quished us, he will leave you unmolested? That wc reason justly, you have sufficient evi- dence before you. If his hostilities were di- rected only against us, with the view of re- venging upon us the former servile condition of his nation, he would immediately have inarched into our country, without at all injur- ing or molesting others; he would have

Click on a lat/long to centre the map there.

Bosphorus	41.125,29.125
Thrace	41.333,26.750
Ister	
Agathyrsi	
Persia	
Maeotis	46.500,36.500
Tanais	47.500,39.500

Example Toponym Resolver: CARMEN for tweets

Java and Python implementations (Dresde et al., 2013)

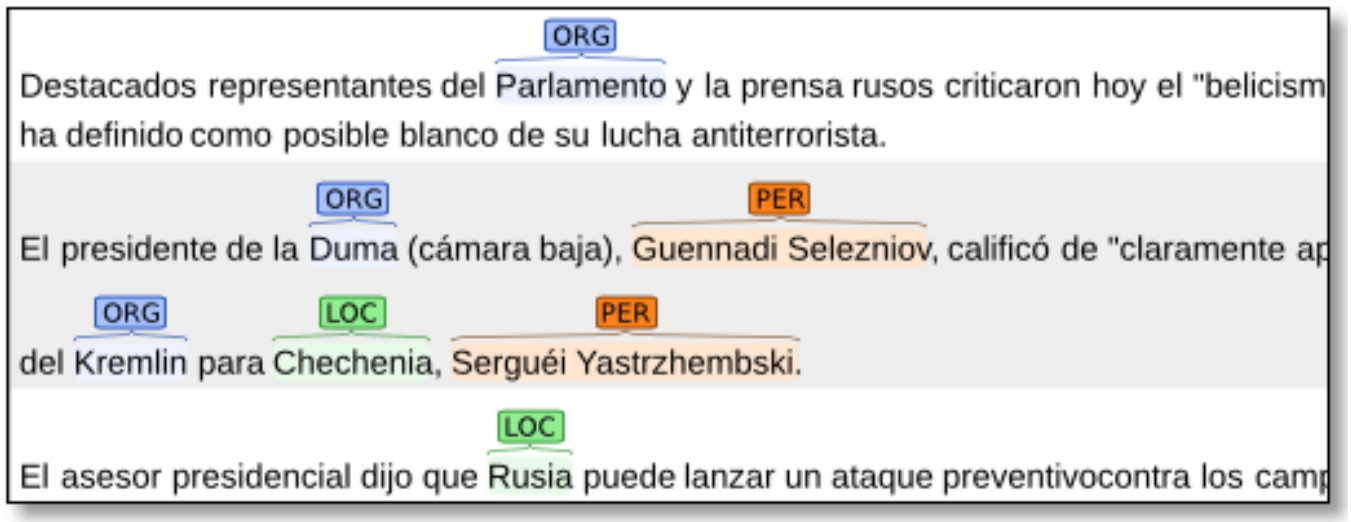
leidner@saturn: ~/carmen-python

File Edit Tabs Help

```
Installed /usr/local/lib/python3.6/dist-packages/geographiclib-1.50-py3.6.egg
Finished processing dependencies for carmen==0.0.4
leidner@saturn:~/carmen-python$ python -m carmen.cli --help
usage: cli.py [-h] [-s] [--order RESOLVERS] [--options OPTIONS]
              [--locations PATH]
              [input_path] [output_path]

Resolve tweet locations.

positional arguments:
  input_path          file containing tweets to locate with geolocation field
                      (defaults to standard input)
  output_path         file to write geolocated tweets to (defaults to standard
                      output)

optional arguments:
  -h, --help          show this help message and exit
  -s, --statistics    show summary statistics
  --order RESOLVERS   preferred resolver order (comma-separated)
  --options OPTIONS   JSON dictionary of resolver options
  --locations PATH    path to alternative location database

Paths ending in ".gz" are treated as gzipped files.
leidner@saturn:~/carmen-python$
```

Annotation & Tools

- MITRE Callisto: commercial open source
- **BRAT** (`http://brat.nlplab.org`):

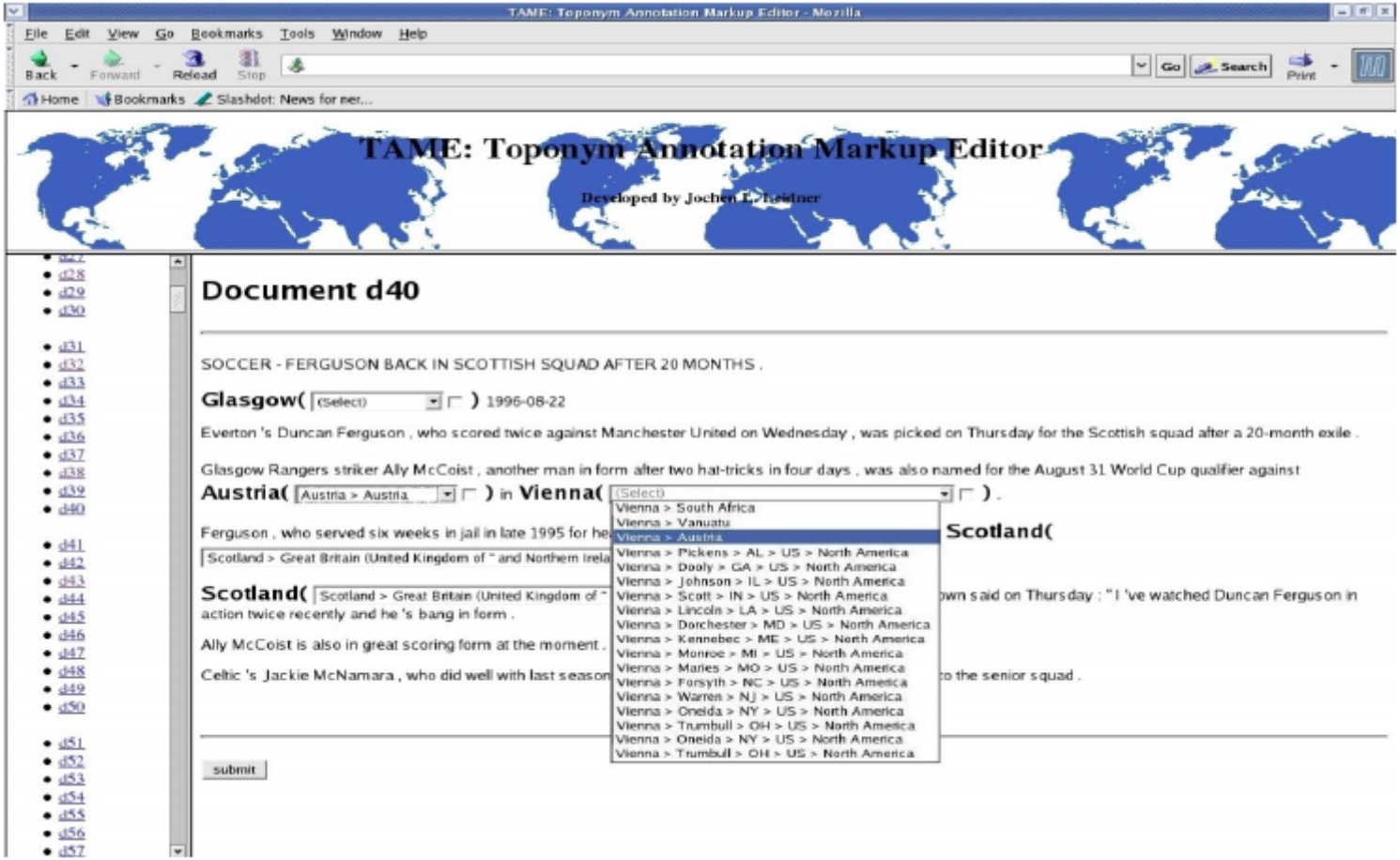


(Source: ibd.) free, Web-based, self-hosted, collaborative

- tagtog (online, `https://www.tagtog.net`)
- Recogito (online `recogito.pelagios.org`, digital humanities focus)

Annotation & Tools (cont'd)

- **TAME** (Leidner, 2007): commercial (contact the author)



- Explosion AI Prodigy: commercial

Text & Meta-Data

- Geographic information that was implicit (encoded in text) can be made explicit as **meta-data**
- Bridge to information retrieval: meta-data can be indexed/searched

Example:

```
<!-- RDF geo meta-data: Berlin (Mitte), Germany -->
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:geo="http://www.w3.org/2003/01/geo/wgs84_pos#">
  <geo:Point>
    <geo:lat>52.531677</geo:lat>
    <geo:long>13.381777</geo:long>
  </geo:Point>
</rdf:RDF>
```

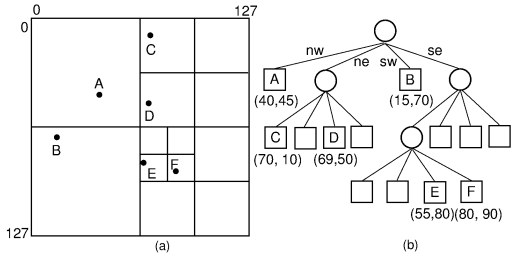
(one of many representations)

Where to Go from Here?

- GIR Workshop Series (Ross Purves/Chris Jones): ⟨`http://www.geo.uzh.ch/~rsp/gir19/index.html`⟩
- UK JISC GEOREFERENCING mailing list (Jochen Leidner): ⟨`https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=GEOREFERENCING`⟩

Spatial Data Structures for Indexing

- Quadtrees
  - quad-trees or Point-Region quadtrees: each node is either a leaf or has exactly 4 child nodes
  - divide 2-dimensional pane into 4 equal quadrants (and further recursively into sub-quadrants)
  - leaf node contains 0-1 single points



- KD-Trees
  - multidimensional extensions of binary search trees
  - efficient processing of search keys across multiple dimensions
  - E.g.: 2-dimensional search using x/y-coordinates (lat/lon)
- Grid Maps
  - divide 2-dimensional plane into $N \times M$ grid elements ("cells")
  - constant-time direct access of cells via coordinates, linear search or hashing inside of cells

Applications

- **Story visualisation**: the generation of a polygon representing the 'spatial aboutness' of a narrative, as a visual spatial summary (e.g. for map focus selection);
- **Spatial browsing**: documents can be explored using spatial dimensions after resolving their toponyms;
- **Answering spatial questions**: given a discourse model for a text that includes resolved toponyms, spatial questions about it can be answered accurately in a knowledge-based fashion;
- **Geographic information retrieval (GIR)**: geo-filters, re-rankers aiming to improve document retrieval quality by taking into account geographic relevance in addition to topic relevance.

Spatial(ly Aware) Question Answering

- Feature type questions: What is X?
- Location questions: Where is X located?
- Routing questions: How can I get to X (from here)? What is on the way (from here) to X?
- Administrative/constituency/partonymy questions: What X is Y part of?
- Distance questions: How far is X from Y?
- Topological questions: What is between X and Y? Which country/city is adjacent to X?
- "Viewshed" questions: What is that over there?
- Event questions: What happened in X in Y?

Open GIR Research Questions

- What efficient, persistent spatial data-structures are not based on B-trees or hashing?
- What are better toponym resolution strategies?
- Which cache replacement strategies are most effective when using spatial access to data?
- How to best integrate geographic (spatial) and textual signals in retrieval?
- What spatial operators may be needed by the various digital humanities disciplines so their research questions can be supported?
- How can we help social scientists to provide them with better proxies than counting sets of keywords to measure certain variables?
- How can we better distinguish place-names invoked via metonymy versus literal ones (e.g. “before Wuhan happened”)

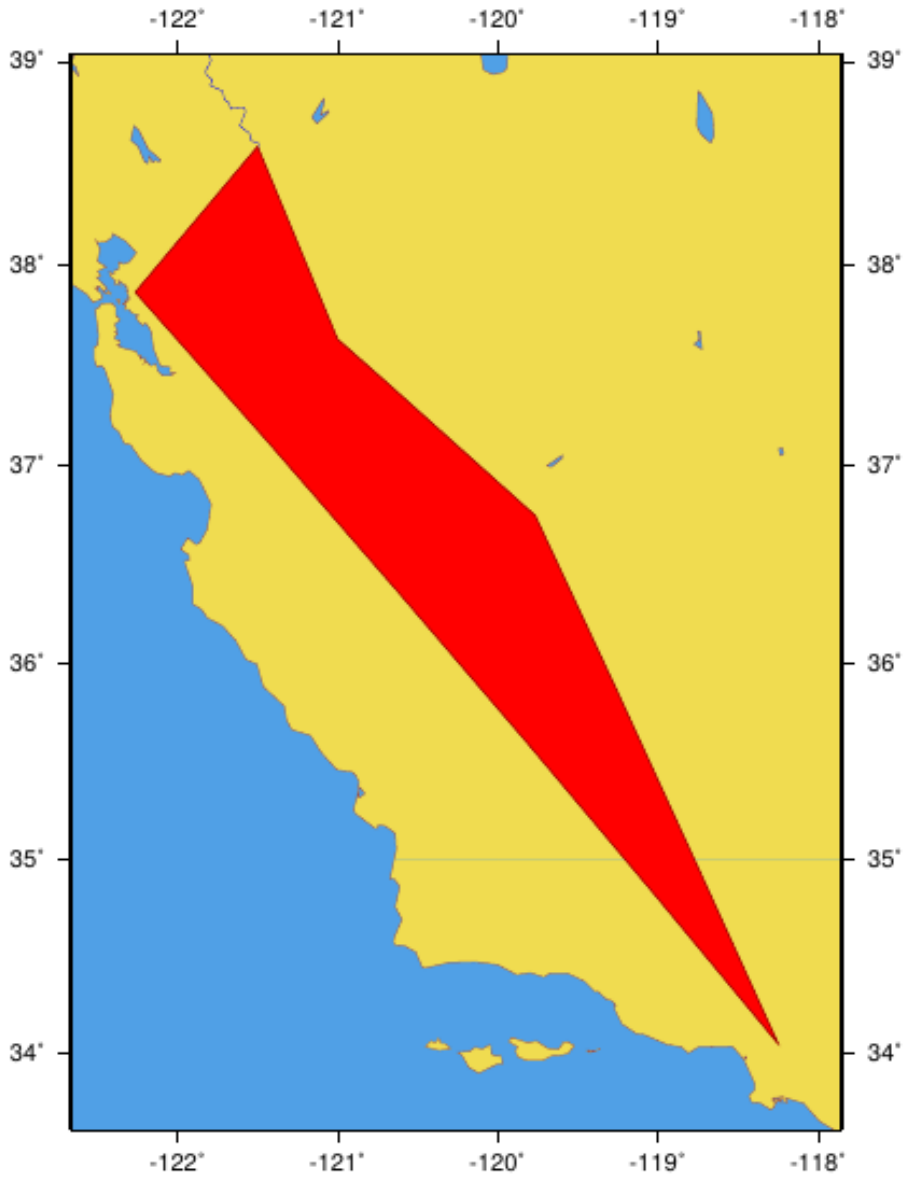
Review Questions

1. What is the difference between location named entity tagging and toponym resolution?
2. What is the difference between geocoding and toponym resolution?
3. What is needed to resolve ambiguous toponyms?
4. Which heuristics have commonly been used to resolve ambiguous toponyms?

Practical Exercises

1. (a) Apply a named entity tagger like OpenCalais or SpaCy to the Samuel Johnson text and conduct an error analysis of 10 randomly chosen pages.
   (b) convert the document to HTML: maintain the original pagination and produce landing pages for each recognized toponym that lists links to all occurrences of the same place name (Hint: a KWIC concordance helps the browsing reader to make sense of the contexts in which the names are mentioned).
2. As a project, create a program that generates a geographic map that depicts the travel trajectory of Samuel Johnson as he travels Scotland.

Missing Woman Map (Leidner et al., 2003)



Putting on Maps with KML (Leidner, 2007

TST4 - MUC4 - 0012 Lima , 14 Aug 88 (AFP) - - [Excerpts] Today it was reported that yesterday the Peruvian police confiscated weapons belonging to the DEA , a U . S . organization engaged in the fight against drug trafficking . The arms were in the possession of three members of an international criminal group called the Green Leaf . [passage omitted] After an intense shoot - out in the Ate - Vitarte district of Lima , personnel of the Anti - Terrorist Directorate have arrested [name indistinct] Reyes , 44 ; Carlos Alonso Rivera , 39 ; and Guillermo Zevallos , 36 . They have been charged with drug trafficking . The police caught the drug traffickers selling drugs on the street in broad daylight . Police confiscated two U . S . - made 45 caliber automatic pistols bearing the DEA acronym and 6 kg of cocaine hydrochloride worth $ 500 , 000 .

(a)

(b)

(c)

References

- E. Amitay et al. 2004. Web-a-Where: Geotagging Web content. *Proc. SIGIR*
- R.S. Beaman and B.J. Conn. 2003. Automated geoparsing and georeferencing of Malesian collection locality data. *Telopea* **10***(1)*
- F. Bilhaut et al. 2003. Geographic reference analysis for geographic document querying, *HLT-NAACL 2003 Workshop: Analysis of Geographic References*
- P. Clough. 2005. Extracting metadata for spatially-aware information retrieval on the Internet, *GIR Workshop held at CIKM*
- P. Clough and M. Sanderson. 2004. A proposal for comparative evaluation of automatic annotation for geo-referenced documents, *GIR Workshop held at SIGIR*
- P. Clough et al. 2004. Extraction of semantic annotations from textual Web pages. *SPIRIT Tech. Rep. D15 6201*, University of Sheffield
- S. Crosier. 2004. *Geocoding in ArcGIS 9*, ESRI Press

References (cont'd)

- M.R. Curry. 1999. Rethinking privacy in a geocoded world. In P.A. Longley et al. eds., *Geographical Information Systems*, vol. 2, chapter 55
- G. DeLozier, J. Baldridge and L. London. 2015. "Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles", *Proc. AAAI*
- M. Dredze et al. 2013. Carmen: A Twitter Geolocation System with Applications to Public Health, Workshop on Expanding the Boundaries of Health Informatics Using AI held at AAAI
- F. Gey et al. 2006. GeoCLEF: The CLEF 2005 cross-language geographic information retrieval track overview. *Proc. CLEF 2005*
- Gittings, B.M. (2012) The Gazetteer for Scotland ⟨`http://www.scottish-places.info`⟩ (online, cited 2020-04-10)

References (cont'd)

- M. Gritta, M.T. Pilehvar and N. Collier. 2018. "Which Melbourne? Augmenting Geocoding with Maps"
- Google, Inc. 2006a. Google Earth, online; cited 2006-06-16 http://earth.google.com
- A.G. Hauptmann and A,M. Olligschlaeger. 1999. Using location information from speech recognition of television news broadcasts, *ECSA ETRW Workshop*
- L.L. Hill. 2000. Core elements of digital gazetteers: placenames, categories, and footprints, *Proc. ECDL*
- L.L. Hill. 2006. *Georeferencing: the Geographic Associations of Information*, MIT Press
- R.R. Larson and P. Frontiera. 2004. Spatial ranking methods for geographic information retrieval (GIR) in digital libraries. *Proc. ECDL*
- R.R. Larson et al. 1995. The Sequoia 2000 electronic repository. *Digital Technical Journal of Digital Equipment Corporation* **7***(3)*

References (cont'd)

- J.L. Leidner. 2006a. An evaluation dataset for the toponym resolution task. *Comp., Env. and Urban Systems* **30***(4)*
- J.L. Leidner. 2006b. Experiments with geo-filtering predicates for information retrieval, *Proc. CLEF 2005*
- J.L. Leidner, G. Sinclair and B. Webber. 2003. Grounding spatial named entities for information extraction and question answering. In *Proc. Workshop on the Analysis of Geographic References held HLT/NAACL*
- J.L. Leidner. 2007. *Toponym Resolution in Text* Ph.D. thesis, University of Edinburgh.
- J.L. Leidner. 2008. *Toponym Resolution in Text*, Irving, CA: Universal Press
- J.L. Leidner. 2017. Georeferencing, in: *International Encyclopedia of Geography: People, The Earth, Environment and Technology*, 14 vols., AAG/Wiley

References (cont'd)

- J.L. Leidner. in print/2020, in: A Survey of Textual Data & Geospatial Technology, M. Werner and Y.Y. Chiang (eds.), *Handbook of Big Geospatial Data*, Springer-Nature
- H. Li et al. 2002. Location normalization for information extraction. *Proc. COLING*
- H. Li et al. 2003. InfoXtract location normalization: a hybrid approach to geographic references in information extraction. *HLT-NAACL 2003 Workshop: Analysis of Geographic References*
- P.A. Longley et al. 2005. *Geographic Information Systems and Science*, 2nd ed.
- K. Markert and M. Nissim. 2002. Towards a corpus annotated for metonymies: the case of location names. *Proc. LREC*
- M. Naaman et al. 2006. Assigning textual names to sets of geographic coordinates. *Comp., Env. and Urban Systems* **30***(4)*
- M. Matei-Chesnoiu. 2015. *Geoparsing Early Modern English Drama*, Palgrave Macmillan

References (cont'd)

- A.M. Olligschlaeger and A.G. Hauptmann. 1999. Multimodal information systems and GIS: The Informedia digital video library. *ESRI User Conf.*
- S.E. Overell and S. Rüger. 2006. Identifying and grounding descriptions of places, *GIR WS held at SIGIR*
- S.E. Overell. 2009. *Geographic Information Retrieval: Classification, Disambiguation and Modelling*, Ph.D. thesis, Imperial College, London.
- B. Pouliquen et al. 2006. Geocoding multilingual texts: Recognition, disambiguation and visualisation. *Proc. LREC*
- H. Samet. 1984. The quadtree and related hierarchical data structures. *ACM Comp. Surv.* **16***(2)*
- M. Sanderson and J. Kohler. 2004. Analyzing geographic queries. *GIR Workshop held at SIGIR*
- F. Schilder et al. 2004. Extracting spatial information: grounding, classifying and linking spatial expressions, *GIR Workshop held at SIGIR*

References (cont'd)

- D.A. Smith and G. Crane. 2001. Disambiguating geographic names in a historical digital library, *Proc. ECDL*
- D.A. Smith and G.Mann. 2003. Bootstrapping toponym classifiers, *HLT-NAACL 2003 Workshop: Analysis of Geographic References*
- M. Speriosu and J. Baldridge. 2013. "Text-driven toponym resolution using indirect supervision" *Proc. ACL*
- B.P. Wing and J. Baldridge. 2011. "Simple supervised document geolocation with geodesic grids" *Proc. ACL-HLT*, 955-964.
- B.P. Wing and J. Baldridge. 2014. "Hierarchical discriminative classification for text-based geolocation" *Proc. EMNLP*.
- A. Woodruff and Chr. Plaunt. 1994. GIPSY: Automated geographic indexing of text documents, *JASIST* **45***(9)*
- W. Zong et al. 2005. On assigning place names to geography related Web pages, *Proc. JCDL*

Abstract

Text Meets Space: Geographic Content Extraction, Resolution and Information Retrieval

J.L. Leidner, B. Martins, K. McDonough and R.S. Purves

In this half-day tutorial, we will review the basic concepts of, methods for, and applications of geographic information retrieval, also showing some possible applications in fields such as the digital humanities. The tutorial is organized in four parts. First we introduce some basic ideas about geography, and demonstrate why text is a powerful way of exploring relevant questions. We then introduce a basic end-to-end pipeline discussing geographic information in documents, spatial and multi-dimensional indexing, and spatial retrieval and spatial filtering. After showing a range of possible applications, we conclude with suggestions for future work in the area.