UNIVERSITY OF CAMBRIDGE

What’s missing in geographic parsing?

Advances and Challenges of Geographic Analysis of Text with Application to Disease Monitoring

13th November 2019

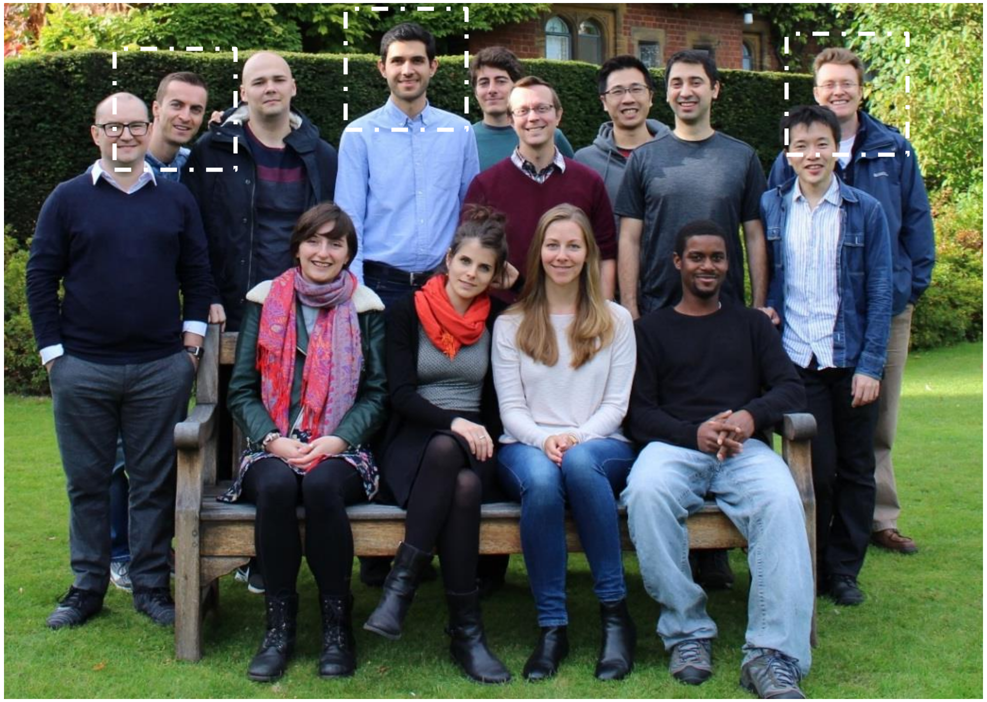
Nigel Collier
nhc30@cam.ac.uk

Language Technology Lab

About the Language Technology Lab

Working on fundamental and applied Natural Language Processing, including:

- Information extraction
- Machine learning
- Machine translation
- Resources and evaluation
- Text generation
- Sentiment analysis
- Social media
- Health applications



Thanks to: Milan Gritta, Taher Pilehvar and Jens Linge (JRC)

Supported by

DREAM DATA, RISK & ENVIRONMENTAL ANALYTICAL METHODS

EPSRC Engineering and Physical Sciences Research Council

MRC Medical Research Council

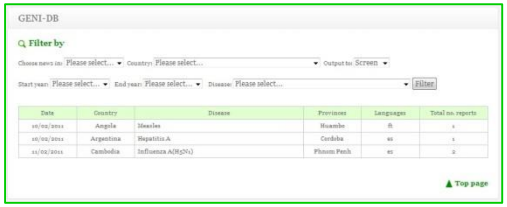
The Alan Turing Institute

Summary of main points

- Epidemic detection from news is a challenging task that will benefit from advances in methods-based research and open source data/software.
- Today we're focussing on Geo-parsing:
  - Geo-parsing is the identification of place names (*toponyms*) in text and their linking to unique identifiers in a databases;
  - Toponym disambiguation *on a global scale at granular levels* is still a great challenge;
  - Need for open standards to compare approaches and involve technical community;
  - Progress with new datasets, neural network models and a taxonomy of toponyms*.*

Experience on epidemic detection with BioCaster (2006-2012)

Event database search



GENI-DB

Filter by

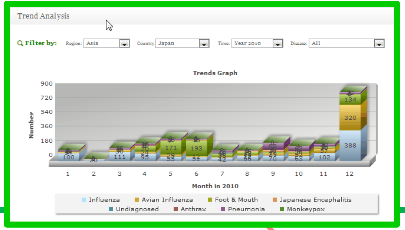
Choose news in: Please select... ▾ Country: Please select... ▾ Output to: Screen ▾

Start year: Please select... ▾ End year: Please select... ▾ Disease: Please select... ▾ [Filter]

Date	Country	Disease	Provinces	Languages	Total no. reports
10/02/2011	Angola	Measles	Huambo	fr	1
10/02/2011	Argentina	Hepatitis A	Cordoba	es	1
11/02/2011	Cambodia	Influenza A(H5N1)	Phnom Penh	es	2

▲ Top page

Trend graphs



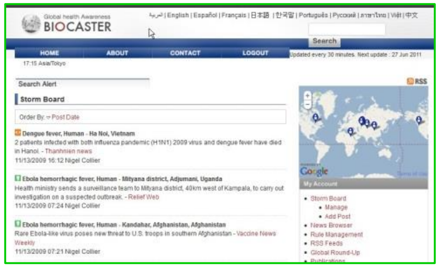
Trend Analysis

Filter by: Region: Asia Country: Japan Time: Year 2010 Disease: All

Ontology browsing

Email/GeoRSS alerting

Watchboard, etc.



Global health Awareness BIOCASTER

العربية | English | Español | Français | 日本語 | 한국말 | Português | Русский | ภาษาไทย | Việt | 中文

Search

HOME ABOUT CONTACT LOGOUT

Updated every 30 minutes. Next update : 27 Jun 2011 17:15 Asia/Tokyo

Search Alert

Storm Board

Order By: ▿ Post Date

Dengue fever, Human - Ha Noi, Vietnam
2 patients infected with both influenza pandemic (H1N1) 2009 virus and dengue fever have died in Hanoi. - Thanhnien news
11/13/2009 16:12 Nigel Collier

Ebola hemorrhagic fever, Human - Mityana district, Adjumani, Uganda
Health ministry sends a surveillance team to Mityana district, 40km west of Kampala, to carry out investigation on a suspected outbreak. - Relief Web
11/13/2009 07:24 Nigel Collier

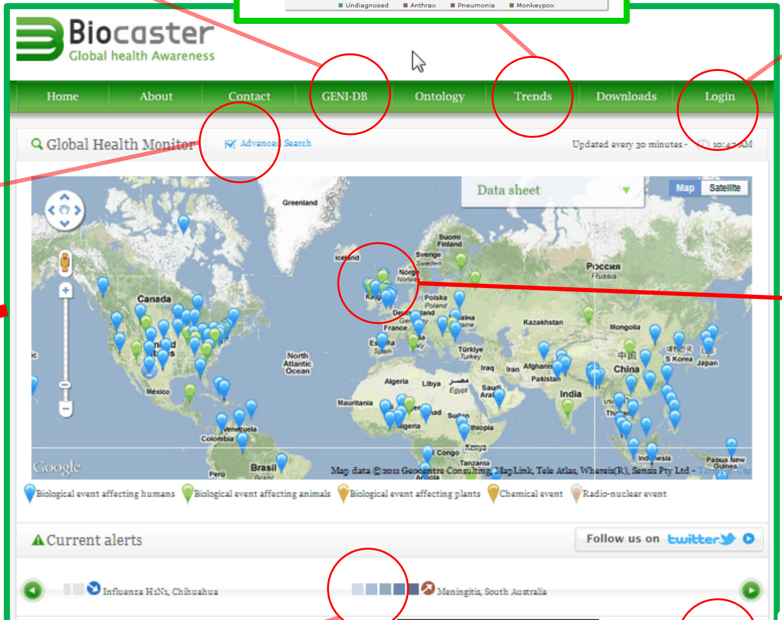
Ebola hemorrhagic fever, Human - Kandahar, Afghanistan, Afghanistan
Rare Ebola-like virus poses new threat to U.S. troops in southern Afghanistan - Vaccine News Weekly
11/13/2009 07:21 Nigel Collier

RSS

My Account

- Storm Board
  - Manage
  - Add Post
- News Browser
- Rule Management
- RSS Feeds
- Global Round-Up
- Publications

Up to date news in 12 languages



Biocaster
Global health Awareness

Home | About | Contact | GENI-DB | Ontology | Trends | Downloads | Login

Global Health Monitor | Advanced Search | Updated every 30 minutes - 10:42AM

Data sheet ▾ | Map | Satellite

Biological event affecting humans | Biological event affecting animals | Biological event affecting plants | Chemical event | Radio-nuclear event

Current alerts | Follow us on twitter

Influenza H1N1, Chihuahua | Meningitis, South Australia

Trend Graph | city trends on Twitter | dizie

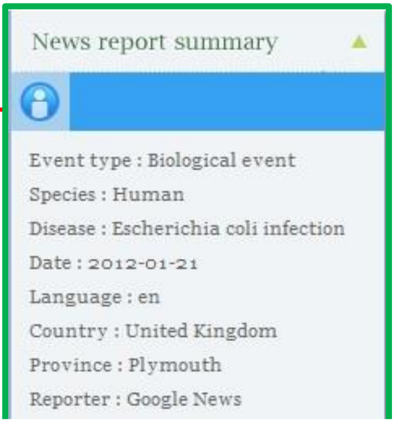
GeoRSS

WHO
EU
IT
JP
CA

GHSI partners

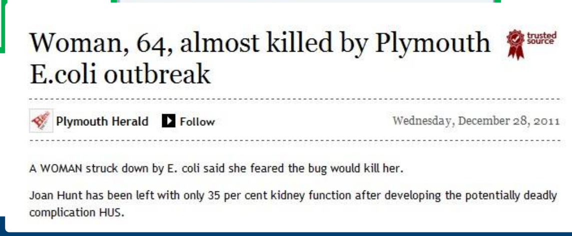
US
UK
FR
DE

Event summaries



News report summary

Event type : Biological event
Species : Human
Disease : Escherichia coli infection
Date : 2012-01-21
Language : en
Country : United Kingdom
Province : Plymouth
Reporter : Google News



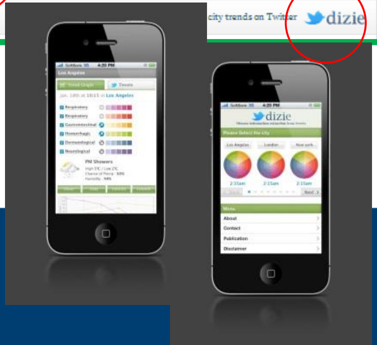
Woman, 64, almost killed by Plymouth E.coli outbreak

trusted source

Plymouth Herald | Follow | Wednesday, December 28, 2011

A WOMAN struck down by E. coli said she feared the bug would kill her.

Joan Hunt has been left with only 35 per cent kidney function after developing the potentially deadly complication HUS.



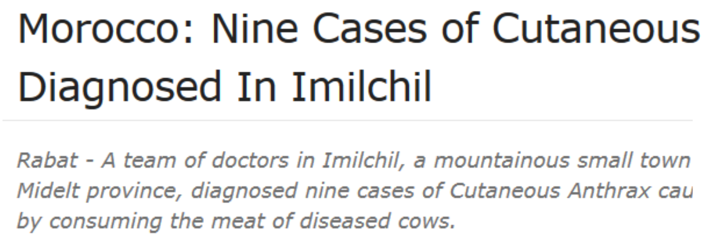
Event alerts

Meningitis, South Australia

Multiple technical challenges raised (2006-2012)

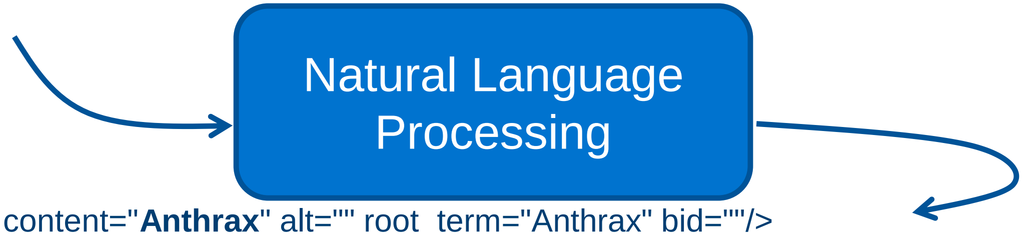
- Geographic parsing
- Trustworthiness of sources (veracity detection)
- Symptom coding (e.g. to ICD-10, SNOMED CT)
- …

Practically speaking there’s no event without time and space



Morocco: Nine Cases of Cutaneous Anthrax Disease Diagnosed In Imilchil

Rabat - A team of doctors in Imilchil, a mountainous small town Midelt province, diagnosed nine cases of Cutaneous Anthrax cau by consuming the meat of diseased cows.



Natural Language Processing

```
<SLOT name="HAS_DISEASE" type="DISEASE" content="Anthrax" alt="" root_term="Anthrax" bid=""/>
    <SLOT name="HAS_LOCATION.COUNTRY" type="LOCATION" content="Morocco" alt="" root_term="Morocco" bid=""/>
    <SLOT name="HAS_LOCATION.PROVINCE" type="LOCATION" content=“Imilchil" alt="" root_term="" bid=""/>
    <SLOT name="HAS_AGENT" type="micro_organism" content="Bacillus anthracis" alt="" root_term="" bid=""/>
    <SLOT name="HAS_SPECIES" type="animal" content="human" alt="" root_term="" bid=""/>
    <SLOT name="TIME.relative" type="string" content=""/>
    <SLOT name="INTERNATIONAL_TRAVEL" type="Boolean" content="false"/>
    <SLOT name="DELIBERATE_RELEASE" type="Boolean" content="false"/>
    <SLOT name="ZOONOSIS" type="Boolean" content="false"/>
    <SLOT name="DRUG_RESISTANCE" type="Boolean" content="false"/>
    <SLOT name="FOOD_CONTAMINATION" type="Boolean" content="false"/>
    <SLOT name="HOSPITAL_WORKER" type="Boolean" content="false"/>
    <SLOT name="FARM_WORKER" type="Boolean" content="false"/>
    <SLOT name="MALFORMED_PRODUCT" type="Boolean" content="false"/>
    <SLOT name="NEW_TYPE_AGENT" type="Boolean" content="false"/>
    <SLOT name="SERVICE_DISRUPTION" type="Boolean" content="false"/>
    <SLOT name="CATEGORY_A" type="Boolean" content="true">
</EVENT>
```

The consequences of getting geoparsing wrong

Equine flu: more horses diagnosed in Camden

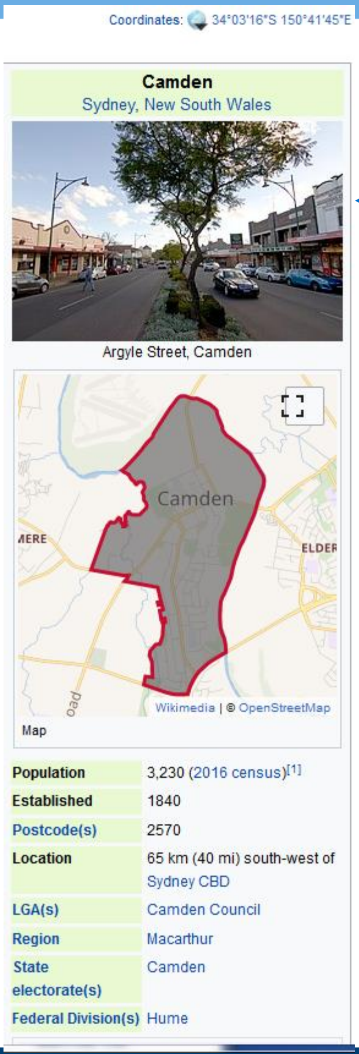
UK ?

VS

Equine flu: more horses diagnosed in Camden

Australia ?

The benefits of getting geoparsing right



Coordinates: 34°03′16″S 150°41′45″E

Camden

Sydney, New South Wales

Argyle Street, Camden

Map

Population	3,230 (2016 census)[1]
Established	1840
Postcode(s)	2570
Location	65 km (40 mi) south-west of Sydney CBD
LGA(s)	Camden Council
Region	Macarthur
State electorate(s)	Camden
Federal Division(s)	Hume

Equine flu: more horses diagnosed in Camden

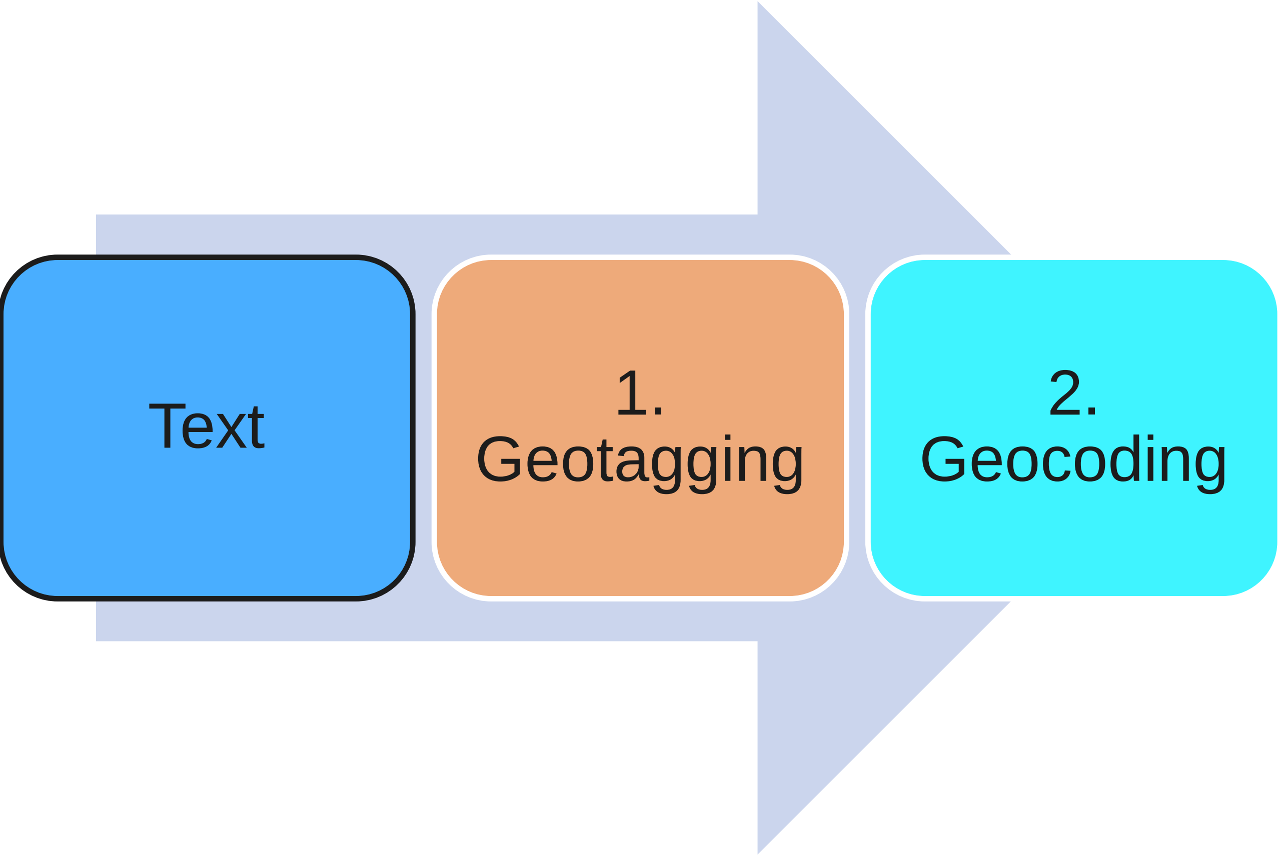
Coordinates:
34° 3′ 16″ S, 150° 41′ 45″ E

Population:
3,230 (2016 census)

Location:
65 km south-west of Sydney

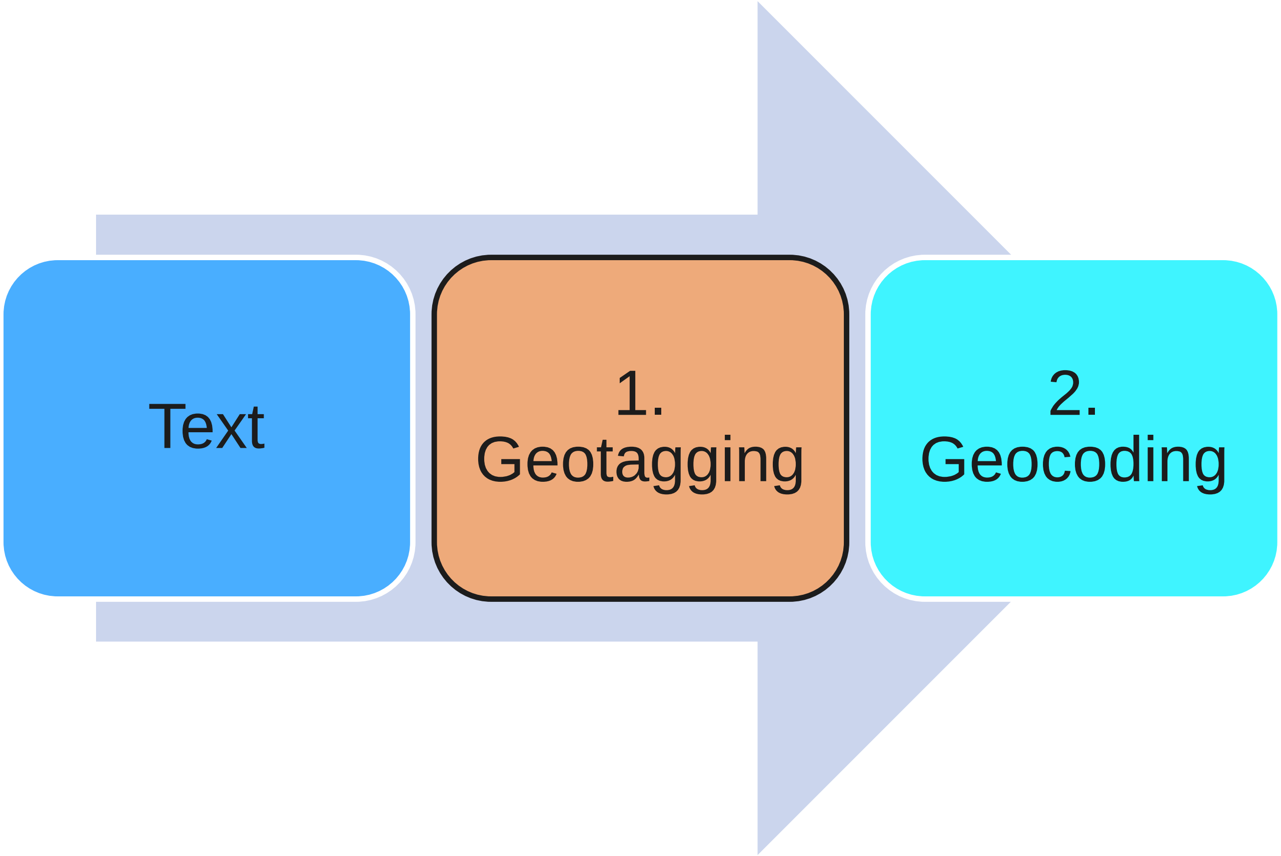
Geoparsing: a two step approach

Equine flu: more horses diagnosed in Camden



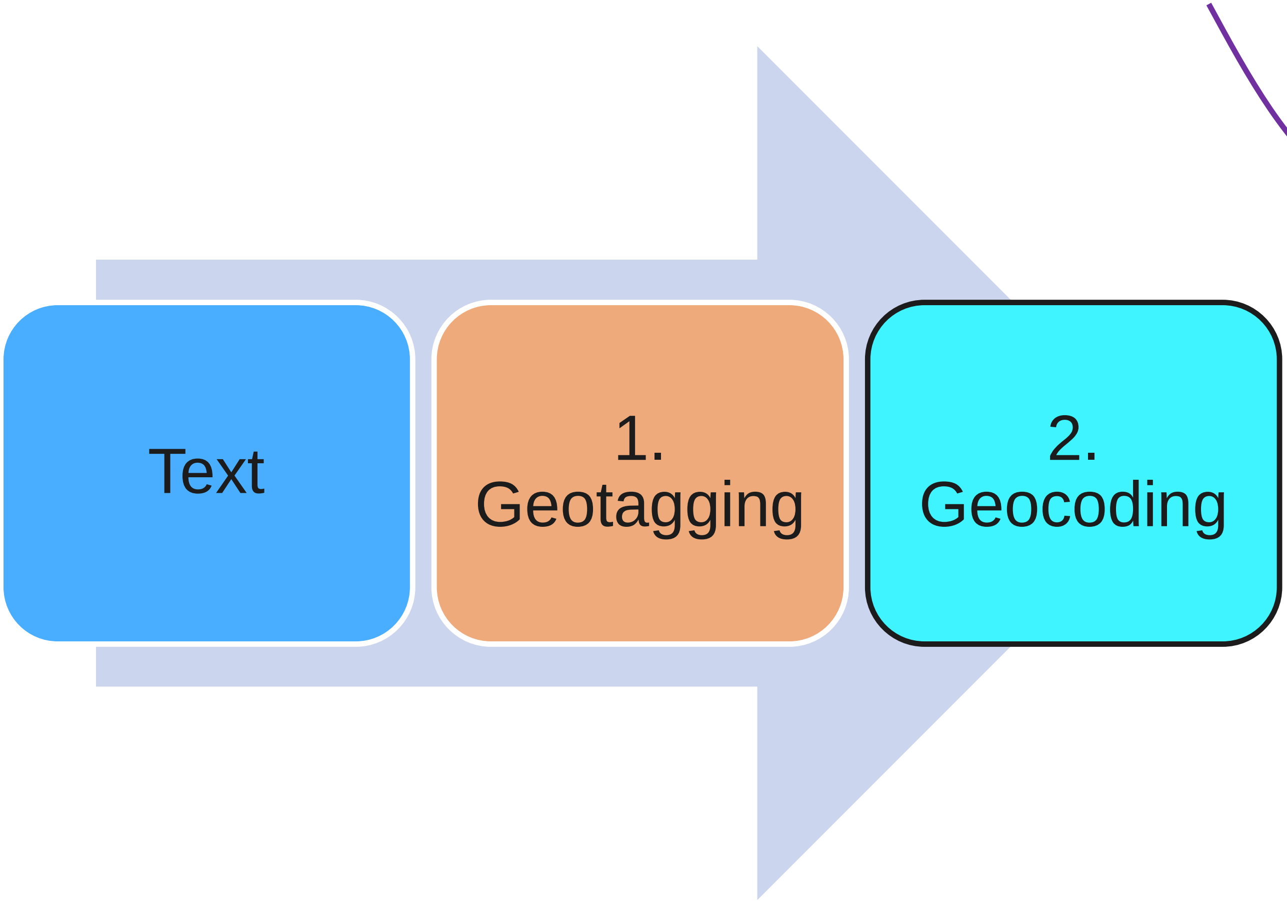
Geoparsing: a two step approach

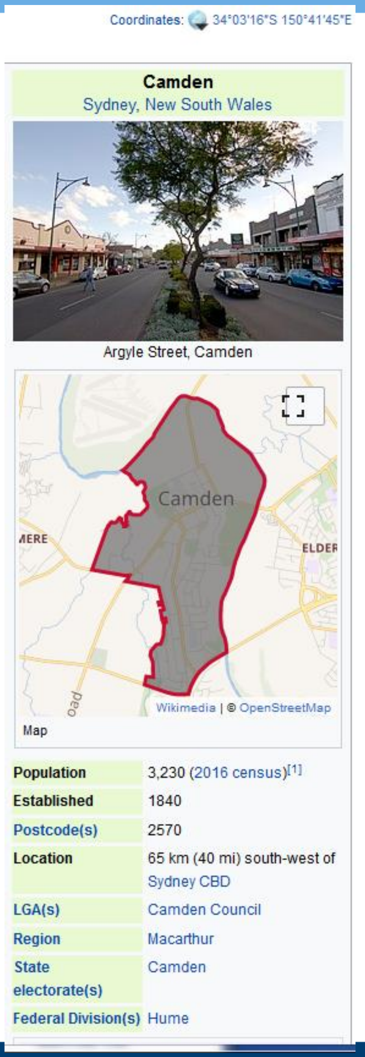
Equine flu: more horses diagnosed in [Camden] LOCATION



Geoparsing: a two step approach

Equine flu: more horses diagnosed in [Camden]$_{34^\circ 3'16''\ \mathrm{S}, 150^\circ 41'45''\mathrm{E}}$





Coordinates: 34°03′16″S 150°41′45″E

Camden

Sydney, New South Wales

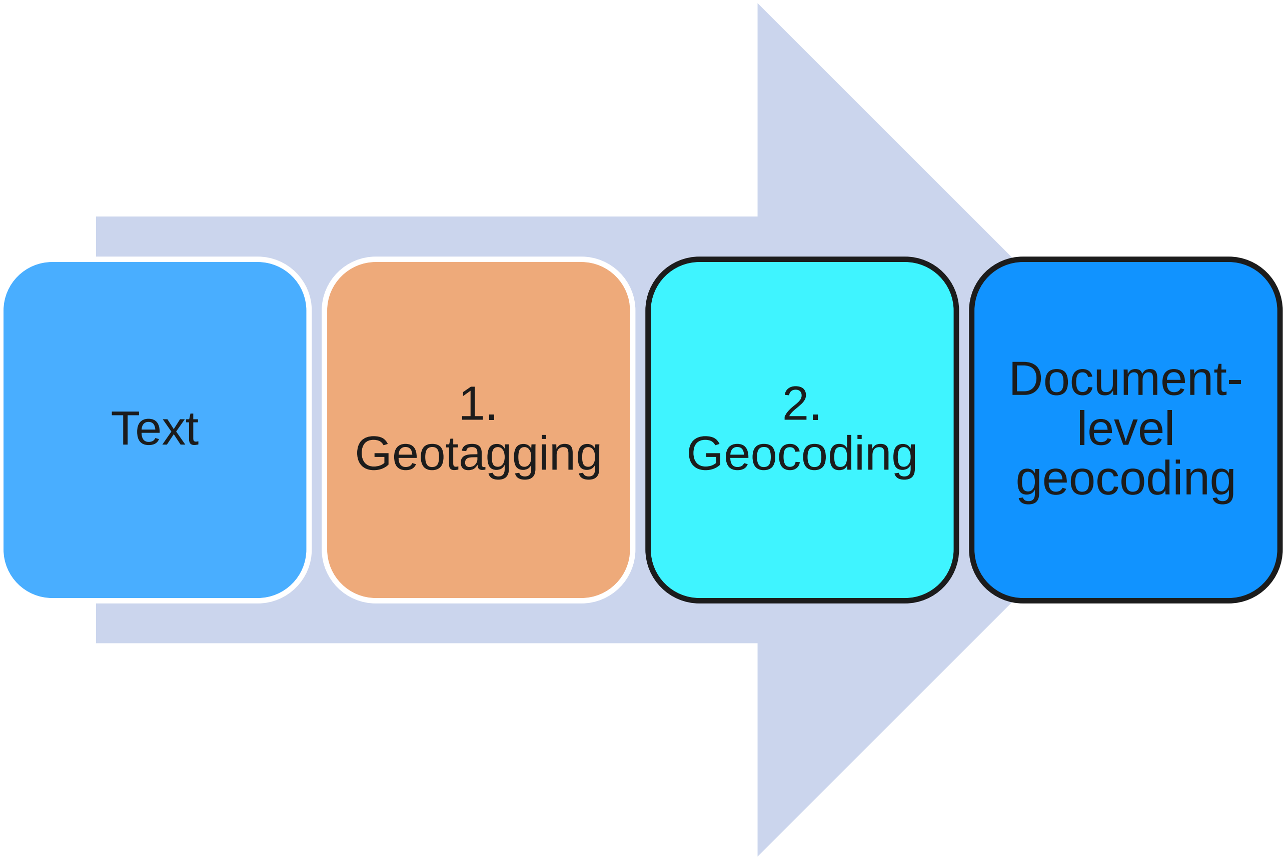
Argyle Street, Camden

Map

Population	3,230 (2016 census)[1]
Established	1840
Postcode(s)	2570
Location	65 km (40 mi) south-west of Sydney CBD
LGA(s)	Camden Council
Region	Macarthur
State electorate(s)	Camden
Federal Division(s)	Hume

Geoparsing: a two step approach

Equine flu: more horses diagnosed in [Camden]$_{34°3′16″ \mathrm{S}, 150°41′45″ \mathrm{E}}$



Text

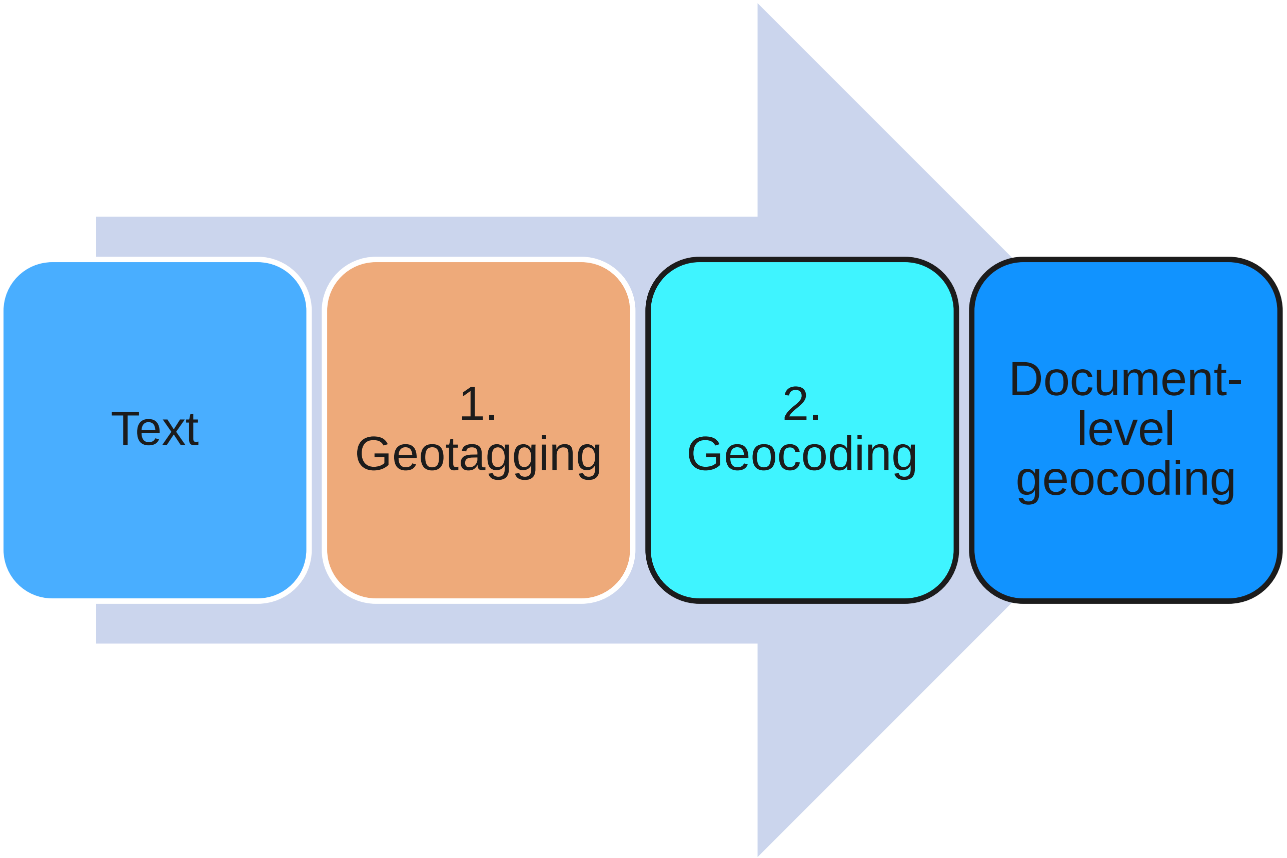
1. Geotagging

2. Geocoding

Document-level geocoding

Geoparsing: a two step approach

Equine flu: more horses diagnosed in [Camden]$_{34°3′16″\text{ S},150°41′45″\text{E}}$



The landscape of geotaggers/coders

Edinburgh geoparser [1] – fully rule-based using local context, spatial clustering and user locality with lists from Wikipedia and Geonames;

CLAVIN [2] – rule based using local context and population priors;

Yahoo! Placemaker [3] – unknown;

GeoTxt [4] – rule-based using local context, approximate string matching and population size;

Topocluster [5] – geo-language model using lexical features;

[1] Grover, C., Tobin, R., Byrne, K., Woollard, M., Reid, J., Dunn, S., & Ball, J. (2010). Use of the Edinburgh geoparser for georeferencing digitized historical collections. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *368*(1925), 3875-3889.
[2] https://clavin.bericotechnologies.com
[3] https://developer.yahoo.com/geo/
[4] Karimzadeh, M., Huang, W., Banerjee, S., Wallgrün, J. O., Hardisty, F., Pezanowski, S., ... & MacEachren, A. M. (2013, November). GeoTxt: a web API to leverage place references in text. In *Proceedings of the 7th workshop on geographic information retrieval* (pp. 72-73). ACM.
[5] DeLozier, G., Baldridge, J., & London, L. (2015, January). Gazetteer-Independent Toponym Resolution Using Geographic Word Profiles. In *AAAI* (pp. 2382-2388).

Rigorous evaluation needs open data standards …

- **War of the Rebellion corpus** [6] - historical texts
- **Wallgrün's Twitter corpus** [7] - tweets
- **TR-CONLL** [8] – news data, proprietary
- **ACE 2005 English SpatialML corpus** – news data, fee-based
- **Local Global Corpus (LGL)** [9] – local news sources around the world

[6] DeLozier, G., Wing, B., Baldridge, J., & Nesbit, S. (2016, August). Creating a novel geolocation corpus from historical texts. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)* (pp. 188-198).
[7] Wallgrün, J. O., Hardisty, F., MacEachren, A. M., Karimzadeh, M., Ju, Y., & Pezanowski, S. (2014, November). Construction and first analysis of a corpus for the evaluation and training of microblog/twitter geoparsers. In *Proceedings of the 8th workshop on geographic information retrieval* (p. 4). ACM.
[8] Leidner, J. L. (2006). An evaluation dataset for the toponym resolution task. *Computers, Environment and Urban Systems*, *30*(4), 400-417.
[9] Lieberman, M. D., Samet, H., & Sankaranarayanan, J. (2010, March). Geotagging with local lexicons to build indexes for textually-specified spatial data. In *2010 IEEE 26th international conference on data engineering (ICDE 2010)* (pp. 201-212). IEEE.

WikToR: a new gold standard corpus

- Designed to test system's ability to disambiguate place names
- 5000 Wikipedia pages containing highly ambiguous place names selected according to the GeoNames database, e.g.
  - Santa Maria (26 entries), Santa Cruz (25 entries), Victoria (23 entries), Lima (19 entries), Santa Barbara (19 entries)
- 200 words for each page to provide context, including the ambiguous place name.
- Ambiguous place names are at least 1000km apart so any mistake by the system is judged to be costly.

[10] Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2018). What's missing in geographical parsing? *Language Resources and Evaluation*, *52*(2), 603-623.

Geotagging performance on the LGL corpus

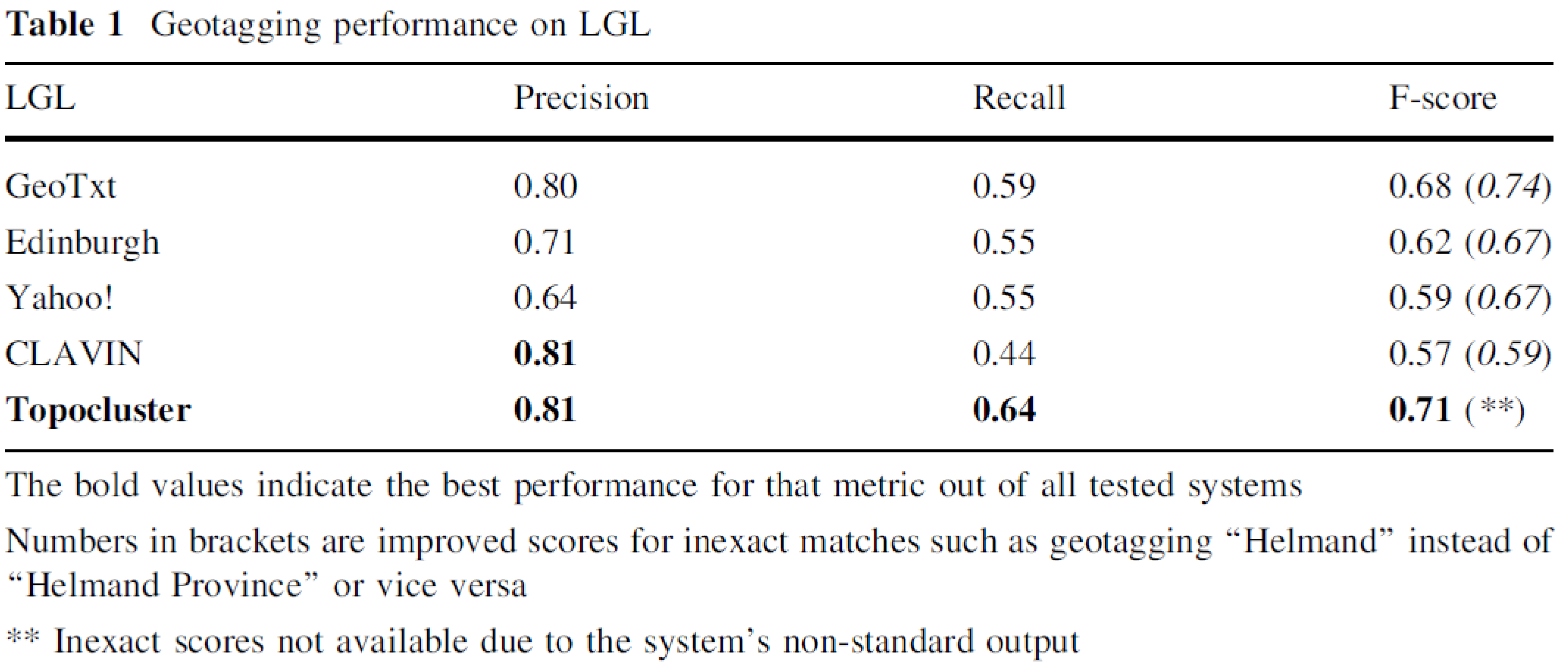


Table 1 Geotagging performance on LGL

LGL	Precision	Recall	F-score
GeoTxt	0.80	0.59	0.68 (*0.74*)
Edinburgh	0.71	0.55	0.62 (*0.67*)
Yahoo!	0.64	0.55	0.59 (*0.67*)
CLAVIN	**0.81**	0.44	0.57 (*0.59*)
Topocluster	**0.81**	**0.64**	**0.71** (**)

The bold values indicate the best performance for that metric out of all tested systems

Numbers in brackets are improved scores for inexact matches such as geotagging “Helmand” instead of “Helmand Province” or vice versa

** Inexact scores not available due to the system’s non-standard output

[10] Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2018). What’s missing in geographical parsing? *Language Resources and Evaluation*, *52*(2), 603-623.

Geocoding performance on the LGL corpus

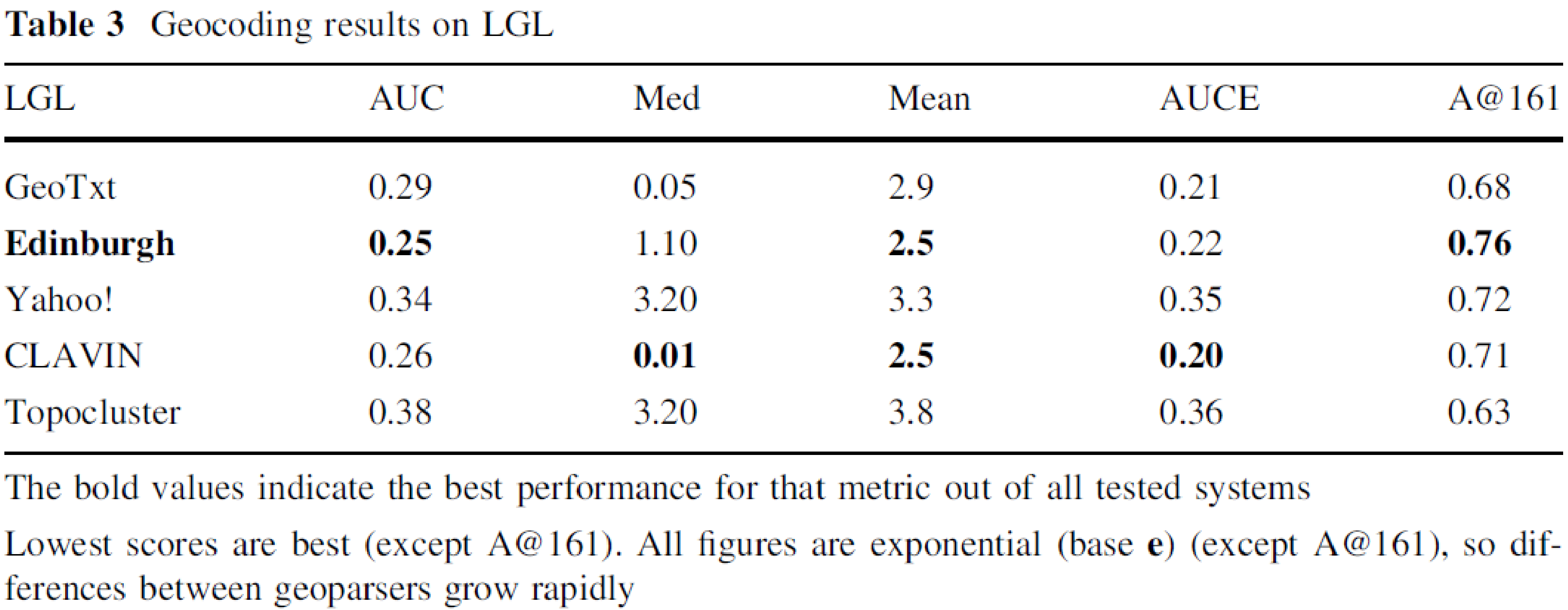


Table 3 Geocoding results on LGL

LGL	AUC	Med	Mean	AUCE	A@161
GeoTxt	0.29	0.05	2.9	0.21	0.68
Edinburgh	**0.25**	1.10	**2.5**	0.22	**0.76**
Yahoo!	0.34	3.20	3.3	0.35	0.72
CLAVIN	0.26	**0.01**	**2.5**	**0.20**	0.71
Topocluster	0.38	3.20	3.8	0.36	0.63

The bold values indicate the best performance for that metric out of all tested systems

Lowest scores are best (except A@161). All figures are exponential (base **e**) (except A@161), so differences between geoparsers grow rapidly

[10] Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2018). What's missing in geographical parsing? *Language Resources and Evaluation*, *52*(2), 603-623.

Geocoding performance on the WikToR corpus

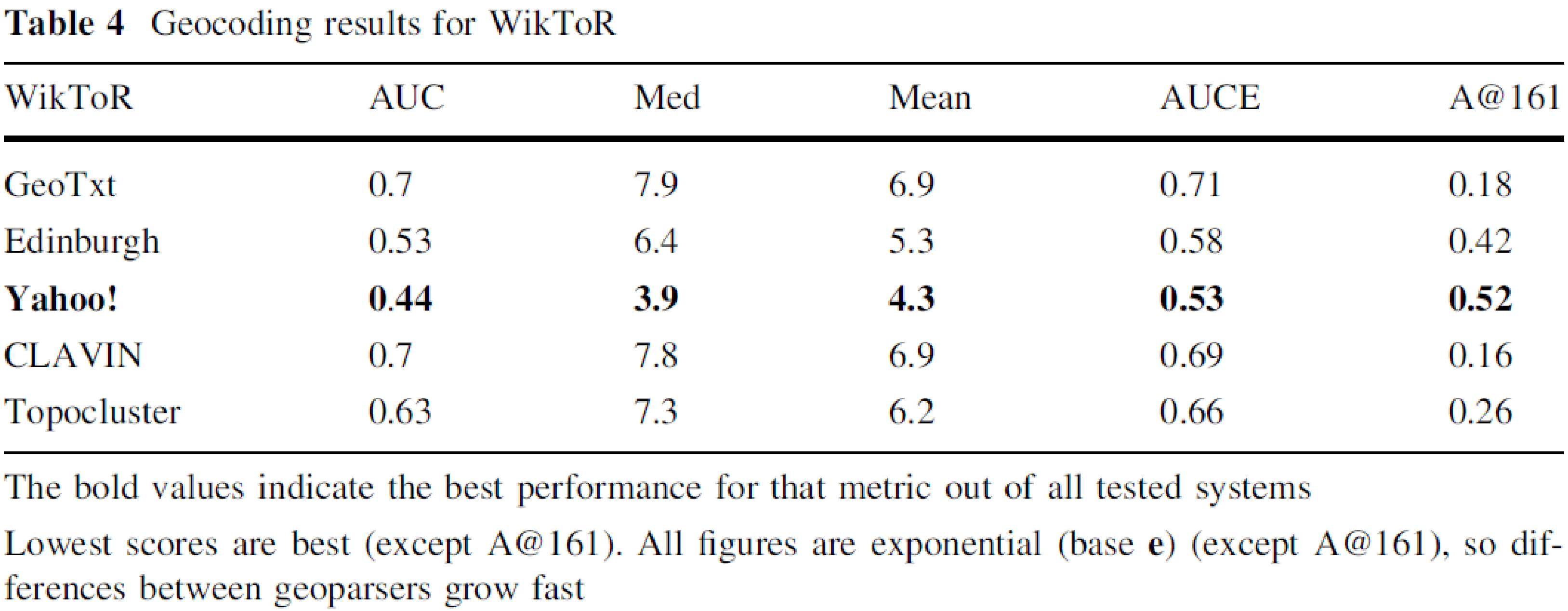


Table 4 Geocoding results for WikToR

WikToR	AUC	Med	Mean	AUCE	A@161
GeoTxt	0.7	7.9	6.9	0.71	0.18
Edinburgh	0.53	6.4	5.3	0.58	0.42
Yahoo!	**0.44**	**3.9**	**4.3**	**0.53**	**0.52**
CLAVIN	0.7	7.8	6.9	0.69	0.16
Topocluster	0.63	7.3	6.2	0.66	0.26

The bold values indicate the best performance for that metric out of all tested systems

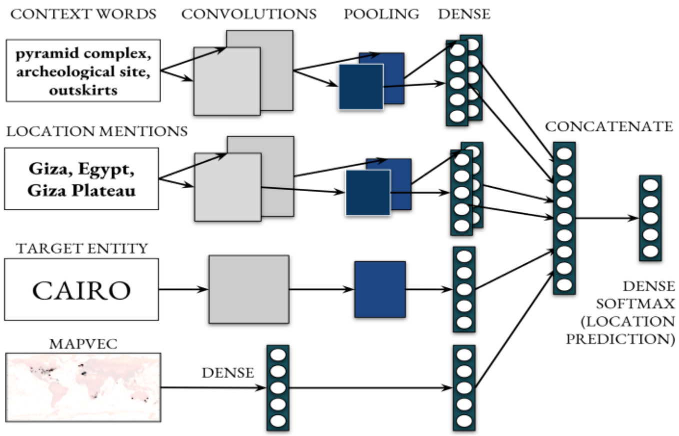
Lowest scores are best (except A@161). All figures are exponential (base **e**) (except A@161), so differences between geoparsers grow fast

[10] Gritta, M., Pilehvar, M. T., Limsopatham, N., & Collier, N. (2018). What's missing in geographical parsing? *Language Resources and Evaluation*, *52*(2), 603-623.

Take homes

- A great geo-parser must excel in
  - Speed (e.g. CLAVIN)
  - Geotagging accuracy (e.g. Topocluster)
  - Geocoding performance (e.g. Yahoo!)
- We’re not there yet.

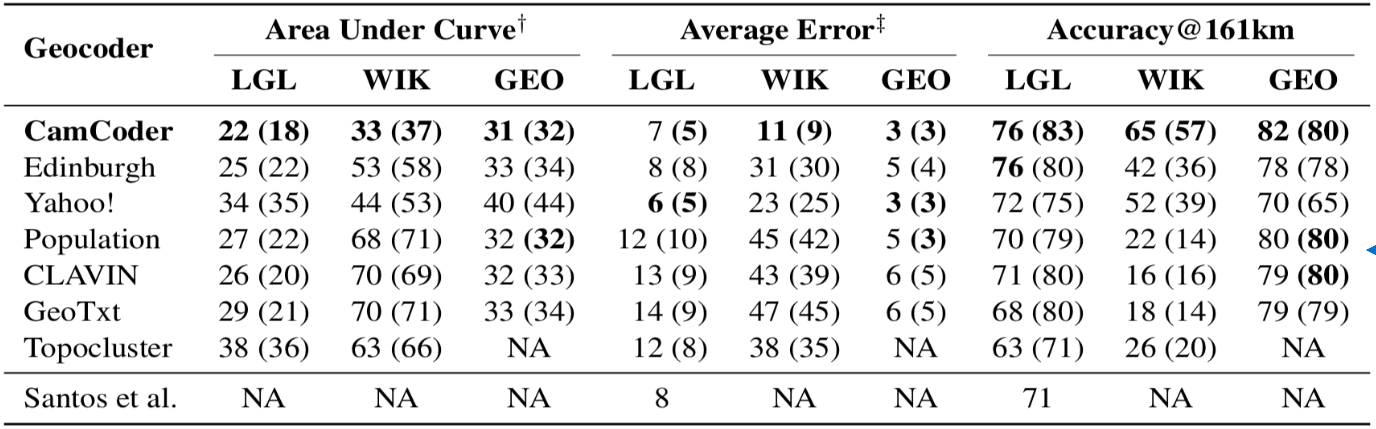
Better geocoding with deep neural networks (CamCoder)



CONTEXT WORDS
CONVOLUTIONS
POOLING
DENSE
pyramid complex, archeological site, outskirts
LOCATION MENTIONS
Giza, Egypt, Giza Plateau
CONCATENATE
TARGET ENTITY
CAIRO
DENSE SOFTMAX (LOCATION PREDICTION)
MAPVEC
DENSE

CamCoder [11]: a state of the art scores on for place name disambiguation on three datasets (Local Global News, WikToR and GeoVirus)

[11] Gritta, M., Pilehvar, M. T., & Collier, N. (2018, July). Which melbourne? augmenting geocoding with maps. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1285-1296).



Geocoder	Area Under Curve†			Average Error‡			Accuracy@161km		
	LGL	WIK	GEO	LGL	WIK	GEO	LGL	WIK	GEO
CamCoder	**22 (18)**	**33 (37)**	**31 (32)**	7 **(5)**	**11 (9)**	**3 (3)**	**76 (83)**	**65 (57)**	**82 (80)**
Edinburgh	25 (22)	53 (58)	33 (34)	8 (8)	31 (30)	5 (4)	**76** (80)	42 (36)	78 (78)
Yahoo!	34 (35)	44 (53)	40 (44)	**6 (5)**	23 (25)	**3 (3)**	72 (75)	52 (39)	70 (65)
Population	27 (22)	68 (71)	32 **(32)**	12 (10)	45 (42)	5 **(3)**	70 (79)	22 (14)	80 **(80)**
CLAVIN	26 (20)	70 (69)	32 (33)	13 (9)	43 (39)	6 (5)	71 (80)	16 (16)	79 **(80)**
GeoTxt	29 (21)	70 (71)	33 (34)	14 (9)	47 (45)	6 (5)	68 (80)	18 (14)	79 (79)
Topocluster	38 (36)	63 (66)	NA	12 (8)	38 (35)	NA	63 (71)	26 (20)	NA
Santos et al.	NA	NA	NA	8	NA	NA	71	NA	NA

A baseline that chooses the most populace location

Drilling down

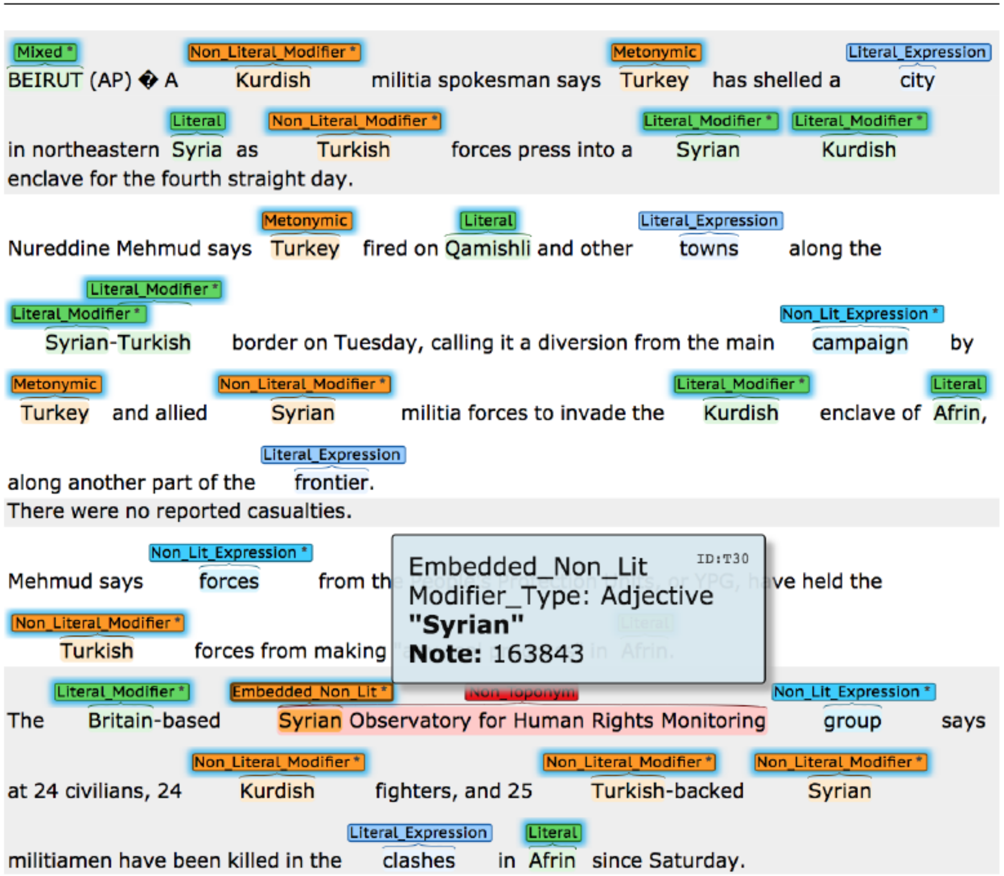
False Positives can be misleading when doing fact extraction:

- **Metonymy** - Moscow negotiates with Seoul.
- **Homonymy** - Milan told me a story.
- **Languages** - She can speak Spanish and Russian.
- **Demonyms** - A Japanese walks into a bar.

False Negatives are often neglected during fact extraction:

- **Coercion** - Meeting is held at the United Nations.
- **Embedded Toponyms** – Athens Festival of Food starts tomorrow.
- **Modifiers** – The target is to reach the Canadian border.

A pragmatic taxonomy of toponyms



BEIRUT (AP) � A Kurdish militia spokesman says Turkey has shelled a city in northeastern Syria as Turkish forces press into a Syrian Kurdish enclave for the fourth straight day.

Nureddine Mehmud says Turkey fired on Qamishli and other towns along the Syrian-Turkish border on Tuesday, calling it a diversion from the main campaign by Turkey and allied Syrian militia forces to invade the Kurdish enclave of Afrin, along another part of the frontier.

There were no reported casualties.

Mehmud says forces from th ve held the

Turkish forces from making Afrin.

Embedded_Non_Lit ID:T30
Modifier_Type: Adjective
"Syrian"
Note: 163843

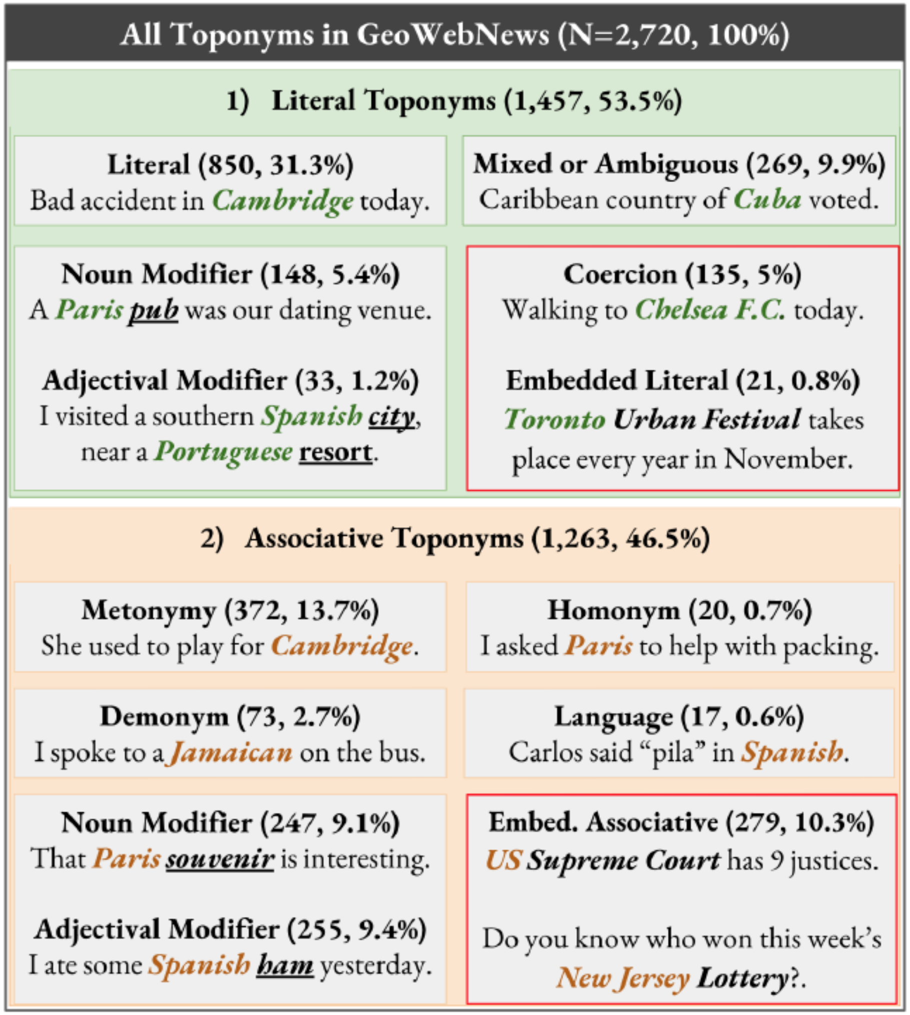
The Britain-based Syrian Observatory for Human Rights Monitoring group says at 24 civilians, 24 Kurdish fighters, and 25 Turkish-backed Syrian militiamen have been killed in the clashes in Afrin since Saturday.

Fig. 4: A GeoWebNews article. An asterisk indicates an attribute, either *a modifier_type* [Adjective, Noun] and/or *a non_locational* [True, False].

Data from the GeoWebNews corpus: 200 news articles from the European Media Monitor

[12] Gritta, M., Pilehvar, M. T. and Collier, N. (2019) "A pragmatic guide to geoparser evaluation" in Language Resources and Evaluation. Published online at https://doi.org/10.1007/s10579-019-09475-3

A pragmatic taxonomy of toponyms



All Toponyms in GeoWebNews (N=2,720, 100%)

1) Literal Toponyms (1,457, 53.5%)

Literal (850, 31.3%)
Bad accident in *Cambridge* today.

Mixed or Ambiguous (269, 9.9%)
Caribbean country of *Cuba* voted.

Noun Modifier (148, 5.4%)
A *Paris* ***pub*** was our dating venue.

Adjectival Modifier (33, 1.2%)
I visited a southern *Spanish* ***city***, near a *Portuguese* **resort**.

Coercion (135, 5%)
Walking to *Chelsea F.C.* today.

Embedded Literal (21, 0.8%)
Toronto ***Urban Festival*** takes place every year in November.

2) Associative Toponyms (1,263, 46.5%)

Metonymy (372, 13.7%)
She used to play for *Cambridge*.

Homonym (20, 0.7%)
I asked *Paris* to help with packing.

Demonym (73, 2.7%)
I spoke to a *Jamaican* on the bus.

Language (17, 0.6%)
Carlos said "pila" in *Spanish*.

Noun Modifier (247, 9.1%)
That *Paris* ***souvenir*** is interesting.

Adjectival Modifier (255, 9.4%)
I ate some *Spanish* ***ham*** yesterday.

Embed. Associative (279, 10.3%)
US ***Supreme Court*** has 9 justices.

Do you know who won this week's *New Jersey* ***Lottery***?.

Data from the GeoWebNews corpus: 200 news articles from the European Media Monitor

[12] Gritta, M., Pilehvar, M. T. and Collier, N. (2019) "A pragmatic guide to geoparser evaluation" in Language Resources and Evaluation. Published online at https://doi.org/10.1007/s10579-019-09475-3

Review

Importance of:

- Methods based research to support epidemic intelligence
- Open source data sets/software for open evaluations and reaching out to technical communities
- Geo-parsing using neural network language models
- Understanding types of toponym mentions

Thank you!

https://sites.google.com/site/nhcollier/

nhc30@cam.ac.uk

ORCID: 0000-0002-7230-4164

Twitter: @nigelhcollier