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
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
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)

- Trend graphs
- Event summaries
- Ontology browsing
- Email/GeoRSS alerting
- Watchboard, etc.

Event database search

Up to date news in 12 languages

WHO
EU
IT
JP
CA

GHSI partners

US
UK
FR
DE

Event alerts

GeoRSS

Experience on epidemic detection with BioCaster (2006-2012)

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

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.
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 in Midelt province, diagnosed nine cases of Cutaneous Anthrax caused by consuming the meat of diseased cows.

<?xml version="1.0" encoding="UTF-8"?>
<Event>
  <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

VS

Equine flu: more horses diagnosed in Camden

UK ?

Australia ?
The benefits of getting geoparsing right

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_

Text

1. Geotagging

2. Geocoding
Geoparsing: a two step approach

Equine flu: more horses diagnosed in [Camden] $34^\circ 3' 16''$ S, $150^\circ 41' 45''$ E
Geoparsing: a two step approach

Equine flu: more horses diagnosed in [Camden]$_{34^\circ3'16''}$ S, $150^\circ41'45''$E
Geoparsing: a two step approach

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


[2] [https://clavin.bericotechnologies.com](https://clavin.bericotechnologies.com)


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


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.

Geotagging performance on the LGL corpus

<table>
<thead>
<tr>
<th></th>
<th>Precision</th>
<th>Recall</th>
<th>F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeoTxt</td>
<td>0.80</td>
<td>0.59</td>
<td>0.68 (0.74)</td>
</tr>
<tr>
<td>Edinburgh</td>
<td>0.71</td>
<td>0.55</td>
<td>0.62 (0.67)</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>0.64</td>
<td>0.55</td>
<td>0.59 (0.67)</td>
</tr>
<tr>
<td>CLAVIN</td>
<td>0.81</td>
<td>0.44</td>
<td>0.57 (0.59)</td>
</tr>
<tr>
<td>Topocluster</td>
<td>0.81</td>
<td>0.64</td>
<td>0.71 (**)</td>
</tr>
</tbody>
</table>

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.

Geocoding performance on the LGL corpus

Table 3  Geocoding results on LGL

<table>
<thead>
<tr>
<th>LGL</th>
<th>AUC</th>
<th>Med</th>
<th>Mean</th>
<th>AUCE</th>
<th>A@161</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeoTxt</td>
<td>0.29</td>
<td>0.05</td>
<td>2.9</td>
<td>0.21</td>
<td>0.68</td>
</tr>
<tr>
<td>Edinburgh</td>
<td><strong>0.25</strong></td>
<td>1.10</td>
<td><strong>2.5</strong></td>
<td>0.22</td>
<td><strong>0.76</strong></td>
</tr>
<tr>
<td>Yahoo!</td>
<td>0.34</td>
<td>3.20</td>
<td>3.3</td>
<td>0.35</td>
<td>0.72</td>
</tr>
<tr>
<td>CLAVIN</td>
<td>0.26</td>
<td><strong>0.01</strong></td>
<td><strong>2.5</strong></td>
<td><strong>0.20</strong></td>
<td>0.71</td>
</tr>
<tr>
<td>Topocluster</td>
<td>0.38</td>
<td>3.20</td>
<td>3.8</td>
<td>0.36</td>
<td>0.63</td>
</tr>
</tbody>
</table>

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.

Geocoding performance on the WikToR corpus

<table>
<thead>
<tr>
<th>WikToR</th>
<th>AUC</th>
<th>Med</th>
<th>Mean</th>
<th>AUCE</th>
<th>A@161</th>
</tr>
</thead>
<tbody>
<tr>
<td>GeoTxt</td>
<td>0.7</td>
<td>7.9</td>
<td>6.9</td>
<td>0.71</td>
<td>0.18</td>
</tr>
<tr>
<td>Edinburgh</td>
<td>0.53</td>
<td>6.4</td>
<td>5.3</td>
<td>0.58</td>
<td>0.42</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>0.44</td>
<td>3.9</td>
<td>4.3</td>
<td>0.53</td>
<td>0.52</td>
</tr>
<tr>
<td>CLAVIN</td>
<td>0.7</td>
<td>7.8</td>
<td>6.9</td>
<td>0.69</td>
<td>0.16</td>
</tr>
<tr>
<td>Topocluster</td>
<td>0.63</td>
<td>7.3</td>
<td>6.2</td>
<td>0.66</td>
<td>0.26</td>
</tr>
</tbody>
</table>

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.

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)

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


<table>
<thead>
<tr>
<th>Geocoder</th>
<th>Area Under Curve†</th>
<th>Average Error‡</th>
<th>Accuracy@161km</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LGL</td>
<td>WIK</td>
<td>GEO</td>
</tr>
<tr>
<td>CamCoder</td>
<td>22 (18)</td>
<td>33 (37)</td>
<td>31 (32)</td>
</tr>
<tr>
<td>Edinburgh</td>
<td>25 (22)</td>
<td>53 (58)</td>
<td>33 (34)</td>
</tr>
<tr>
<td>Yahoo!</td>
<td>34 (35)</td>
<td>44 (53)</td>
<td>40 (44)</td>
</tr>
<tr>
<td>Population</td>
<td>27 (22)</td>
<td>68 (71)</td>
<td>32 (32)</td>
</tr>
<tr>
<td>CLAVIN</td>
<td>26 (20)</td>
<td>70 (69)</td>
<td>32 (33)</td>
</tr>
<tr>
<td>GeoTxt</td>
<td>29 (21)</td>
<td>70 (71)</td>
<td>33 (34)</td>
</tr>
<tr>
<td>Topocluster</td>
<td>38 (36)</td>
<td>63 (66)</td>
<td>NA</td>
</tr>
<tr>
<td>Santos et al.</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>
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

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

<table>
<thead>
<tr>
<th>BEIRUT (AP)</th>
<th>A</th>
<th>Kurdish</th>
<th>militia spokesman says</th>
<th>Turkey</th>
<th>has shelled a</th>
<th>city</th>
</tr>
</thead>
<tbody>
<tr>
<td>in northeastern Syria as Turkish forces press into a Syrian Kurdish enclave for the fourth straight day.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>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.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>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.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

# A pragmatic taxonomy of toponyms

<table>
<thead>
<tr>
<th>All Toponyms in GeoWebNews (N=2,720, 100%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Literal Toponyms (1,457, 53.5%)</td>
</tr>
<tr>
<td>- Literal (850, 31.3%)</td>
</tr>
<tr>
<td>Bad accident in <em>Cambridge</em> today.</td>
</tr>
<tr>
<td>- Mixed or Ambiguous (269, 9.9%)</td>
</tr>
<tr>
<td>Caribbean country of <em>Cuba</em> voted.</td>
</tr>
<tr>
<td>- Noun Modifier (148, 5.4%)</td>
</tr>
<tr>
<td>A <em>Paris pub</em> was our dating venue.</td>
</tr>
<tr>
<td>- Adjectival Modifier (33, 1.2%)</td>
</tr>
<tr>
<td>I visited a southern <em>Spanish city</em> near a <em>Portuguese resort</em>.</td>
</tr>
<tr>
<td>- Coercion (135, 5%)</td>
</tr>
<tr>
<td>Walking to <em>Chelsea F.C.</em> today.</td>
</tr>
<tr>
<td>- Embedded Literal (21, 0.8%)</td>
</tr>
<tr>
<td><em>Toronto Urban Festival</em> takes place every year in November.</td>
</tr>
<tr>
<td>2) Associative Toponyms (1,263, 46.5%)</td>
</tr>
<tr>
<td>- Metonymy (372, 13.7%)</td>
</tr>
<tr>
<td>She used to play for <em>Cambridge</em>.</td>
</tr>
<tr>
<td>- Homonym (20, 0.7%)</td>
</tr>
<tr>
<td>I asked <em>Paris</em> to help with packing.</td>
</tr>
<tr>
<td>- Demonym (73, 2.7%)</td>
</tr>
<tr>
<td>I spoke to a <em>Jamaican</em> on the bus.</td>
</tr>
<tr>
<td>- Language (17, 0.6%)</td>
</tr>
<tr>
<td>Carlos said “pila” in <em>Spanish</em>.</td>
</tr>
<tr>
<td>- Noun Modifier (247, 9.1%)</td>
</tr>
<tr>
<td>That <em>Paris souvenir</em> is interesting.</td>
</tr>
<tr>
<td>- Embed. Associative (279, 10.3%)</td>
</tr>
<tr>
<td><em>US Supreme Court</em> has 9 justices.</td>
</tr>
<tr>
<td>- Adjectival Modifier (255, 9.4%)</td>
</tr>
<tr>
<td>I ate some <em>Spanish ham</em> yesterday.</td>
</tr>
</tbody>
</table>

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