Which Melbourne?
Augmenting Geocoding with Maps
Milan Gritta, Mohammad Taher Pilehvar, Nigel Collier
GOAL: Geolocation of text.

Geoparsing Pipeline:

Reference Geocoding

NER GEOTAGGING

Document Geocoding

NEL + WSD GEOCODING

Geocoding or Toponym Resolution

BREAKING NEWS!!!

Incident outside Melbourne’s Sully’s Backstreet Bar. Suspect taken to the Brevard County Jail.
BREAKING NEWS!!!

Incident outside Melbourne’s Sully’s Backstreet Bar. Suspect taken to the Brevard County Jail.
Geoparsing Systems

RULE-BASED SYSTEMS
- CLAVIN (Open-Source, v2016) – NER
- Edinburgh Parser (Grover et al. 2010) v2016
- GeoTxt (Karimzadeh et al. 2013) v2016 – NER
- Population Baseline (choose highest population)

STATISTICAL (& CLOSED SOURCE)
- Topocluster (De Lozier et al. 2015) v2016 - NER
- Yahoo! Placemaker (Proprietary Algorithm)

MACHINE (DEEP) LEARNING
- LambdaMART (Santos et al. 2015) (no source)
- CamCoder (Gritta et al. 2018) v2018 - NER

Geocoding similar to WSD but...
- Ambiguity of toponyms greater (e.g. 10+ Melbournes in the world)
- Contextual clues not adequate or missing for small (local) places
- Often difficult for humans to judge
- 50% - 75% resolved by population
1 TEXT DOCUMENT, 2 SETS OF FEATURES

The Map Vector

(reshape to) 1D Map Vector

Bag of locations.

Bag of words.

| BAR  | SUSPECT | OUTSIDE | PLACE | INCIDENT | SULLY'S | TAKEN |
|------|---------|---------|-------|----------|---------|-------|
| 0.5  | 0.6     | 1.0     | 0.1   | 0.4      | 0       | 0.2   | 0.9   |

Lexical Footprint

Geographic Footprint

180 DEGREES

360 DEGREES

LONGITUDE

LATITUDE
The Giza pyramid complex is an archaeological site on the Giza Plateau, on the outskirts of Cairo, Egypt.
Evaluation

Datasets

- LOCAL GLOBAL LEXICON (LGL) by (Lieberman et al. 2010) – packaged with our code.
- 588 local news articles from global sources
- 4460 annotated places, Medium Difficulty Test
- WIKIPEDIA TOPONYM RETRIEVAL (WikToR) by (Gritta et al. 2017) – also packaged with our code.
- Wikipedia-based geoparsing of 5,000 articles
- High Difficulty Test, 25,000+ locations in total

Other corpora available (De Lozier et al. 2010), (Wallgrun et al. 2017), (Buscaldi and Rosso 2008), (De Oliveira et al. 2017), (Mani et al. 2010), (Eisenstein et al. 2010) but issues with cost, scope, annotation, size, type of task, completeness, etc.

OR RESOURCES NOT PUBLISHED WITH PAPER
GeoVirus.xml

New Dataset

- 229 articles (August, September 2017)
- NER/Geotagging and Geocoding
- KEYWORDS: Ebola, Bird Flu, Swine Flu, AIDS, Mad Cow Disease, many more. (Medisys JRC)
- Locations: 2,167, Word Count: 63,205
- https://github.com/milangritta

DOWNLOAD
## OVERALL PERFORMANCE COMPARISON

| Geocoder       | System configuration                      | Dataset         | Average |
|----------------|-------------------------------------------|-----------------|---------|
|                |                                           | LGL | WIK | GEO |         |
| Geocoder       | Language Features + MapVec Features |                 |       |
| CamCoder       | CNN + MLP                                | 0.22 | 0.33 | 0.31 | 0.29    |
| Lexical Only   | CNN                                       | 0.23 | 0.39 | 0.33 | 0.32    |
| MapVec Only    | — + MLP                                   | 0.25 | 0.41 | 0.32 | 0.33    |
| Context2vec†   | LSTM + MLP                                | 0.24 | 0.38 | 0.33 | 0.32    |
| Context2vec    | LSTM + —                                  | 0.27 | 0.47 | 0.39 | 0.38    |
| Random Forest  | MapVec features only, no lexical input    | 0.26 | 0.36 | 0.33 | 0.32    |
| Naive Bayes    | MapVec features only, no lexical input    | 0.28 | 0.56 | 0.36 | 0.40    |
| Population     | —                                         | 0.27 | 0.68 | 0.32 | 0.42    |

| Geocoder       | Dataset         | Average |
|----------------|-----------------|---------|
| Context2vec†   |                 |         |
| Context2vec    |                 | 0.27 | 0.47 | 0.39 | 0.38    |
| Random Forest  | MapVec features only, no lexical input    | 0.26 | 0.36 | 0.33 | 0.32    |
| Naive Bayes    | MapVec features only, no lexical input    | 0.28 | 0.56 | 0.36 | 0.40    |
| Population     |                 | 0.27 | 0.68 | 0.32 | 0.42    |
Summary of Contributions

- Lexical CNN
- Geocoder
- The Map Vector
- GeoVirus.xml

**CamCoder**

**Lexical CNN Geocoder**
Money enables much Scientific Research.

Thank You!
THANK YOU and CHECK OUT THE PAPER.
https://github.com/milangritta

Questions?