Language Technology Lab, University of Cambridge

Which Melbourne? Augmenting Geocoding with Maps

Milan Gritta, Mohammad Taher Pilehvar, Nigel Collier

UNIVERSITY OF CAMBRIDGE

BREAKING NEWS!!!

Incident outside Melbourne's Sully's Backstreet Bar. Suspect taken to the Brevard County Jail.

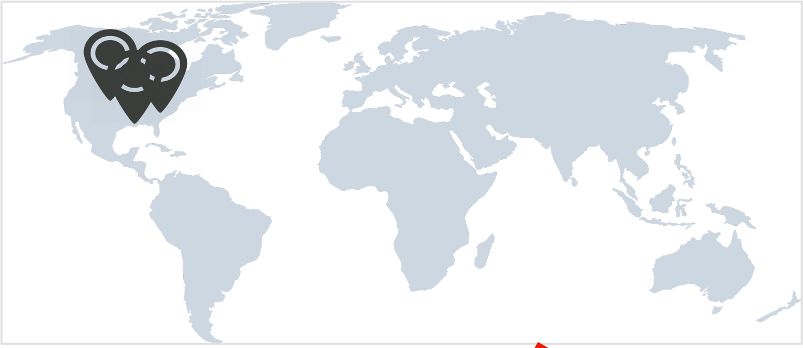
GOAL: Geolocation of text.

Geoparsing Pipeline:

NER
GEOTAGGING

NEL + WSD
GEOCODING

Reference Geocoding

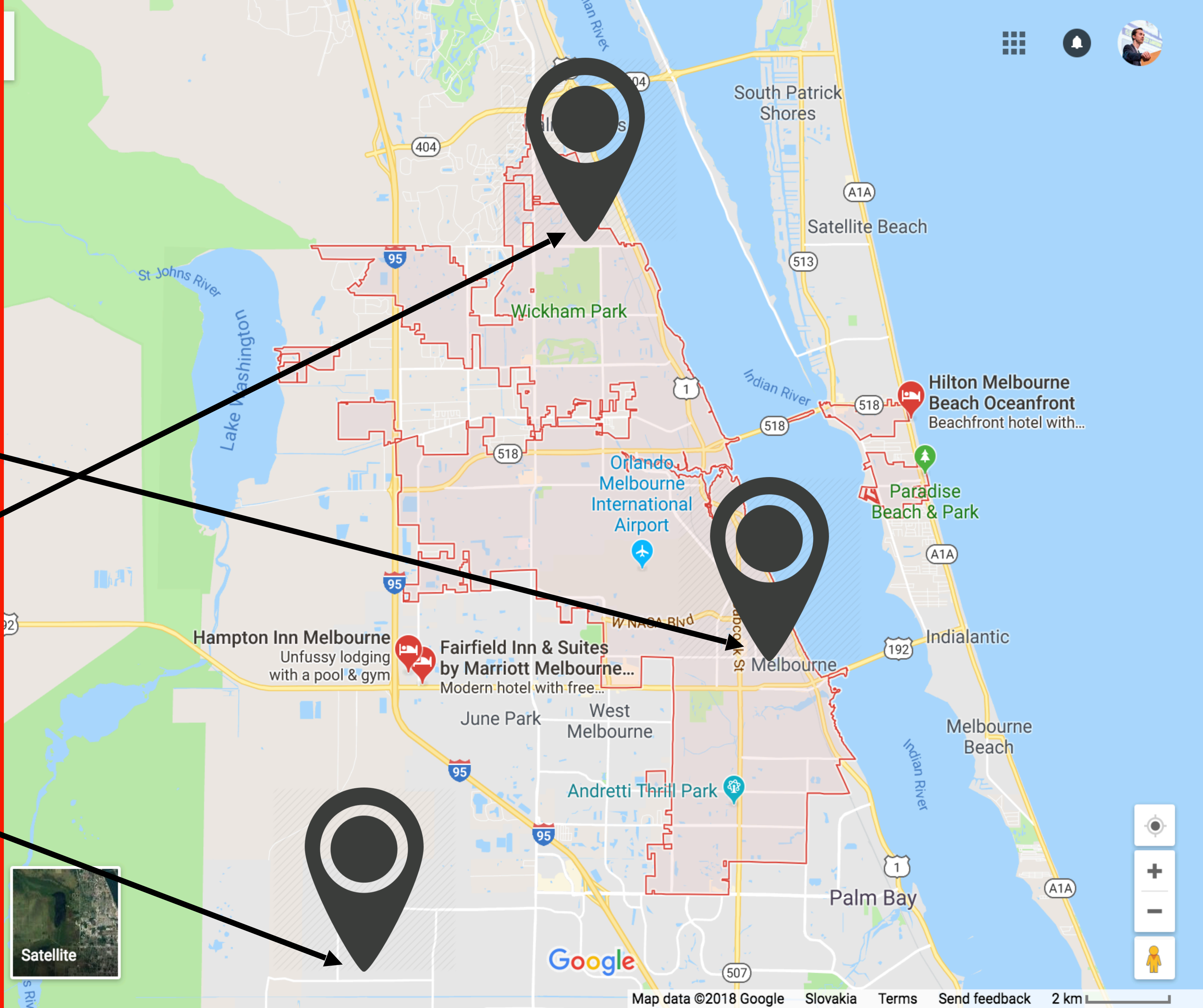


Document Geocoding

Geocoding or Toponym Resolution

BREAKING NEWS!!!

Incident outside Melbourne's Sully's Backstreet Bar. Suspect taken to the Brevard County Jail.



Background

Geoparsing Systems

Geocoding similar to WSD but…

- Ambiguity of toponyms greater (e.g. 10+ Melbournes in the world)
- Contextual clues not adequate :R 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

Lexical Footprint

Geographic Footprint

Bag of words.

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

LONGITUDE

LATTITUDE

0	0	0	0	0	0	0	0
0	0.8	0	**0.2**	1.0	0.6	0	0
0	**0.2**	0.3	0	0	0	0.9	0
0	0.1	0	0	0.1	0	0	0
0	0	0.4	0	0.9	0.5	**0.6**	0
0	0	0	0	0	0	0	0

180 DEGREES

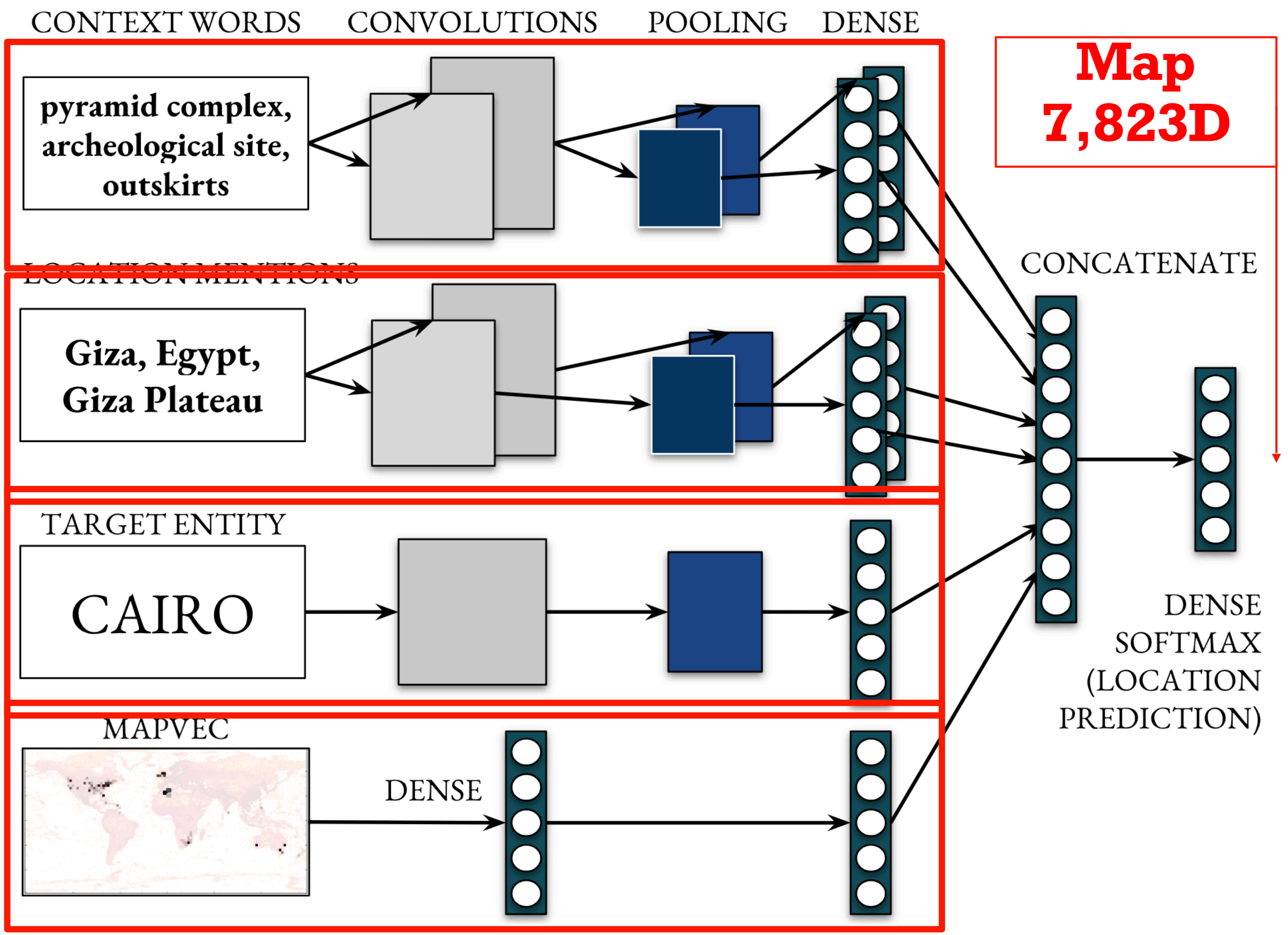
360 DEGREES

(reshape to) 1D Map Vector

Bag of locations.

ARTICLE.COM

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


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

Map 7,823D

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

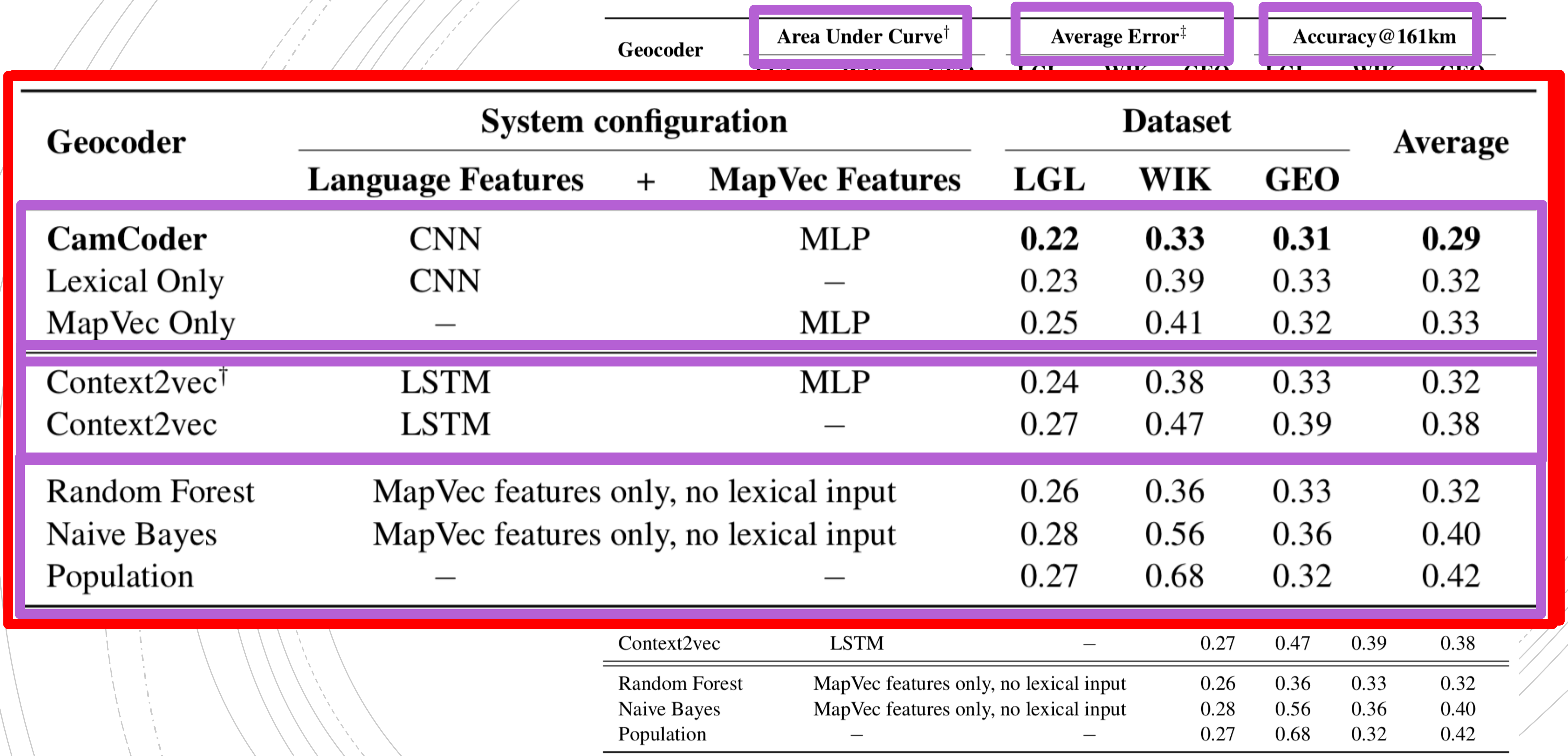
DOWNLOAD

Free as in Freedom

WIKINEWS

- 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**

OVERALL PERFORMANCE COMPARISON

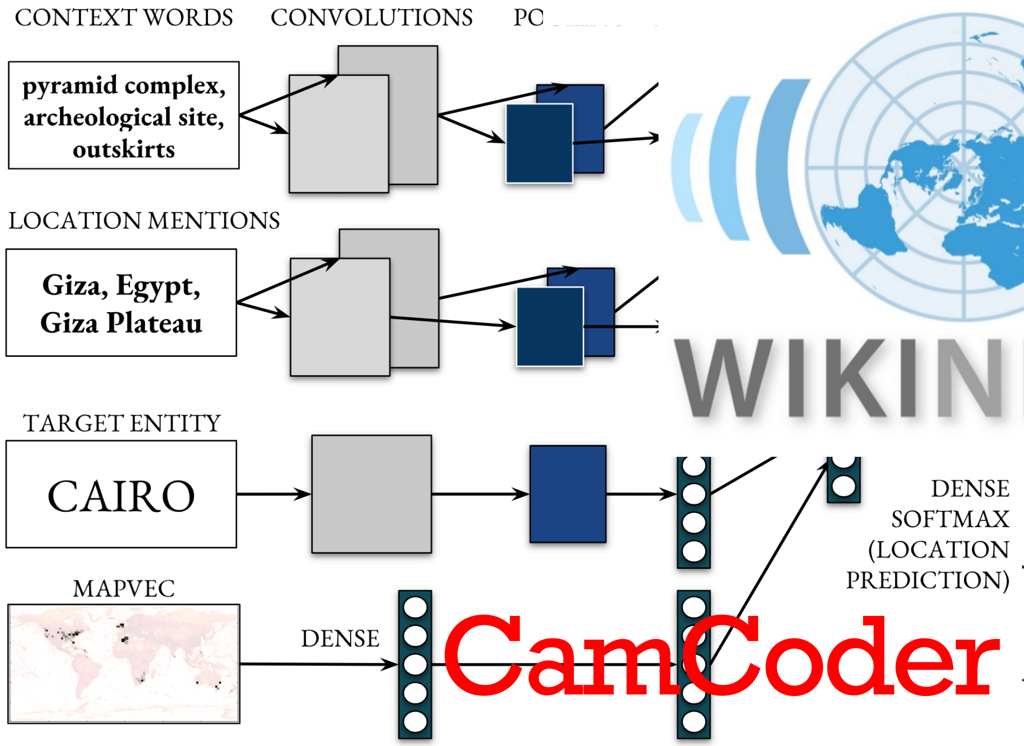


Geocoder	Area Under Curve†	Average Error‡	Accuracy@161km

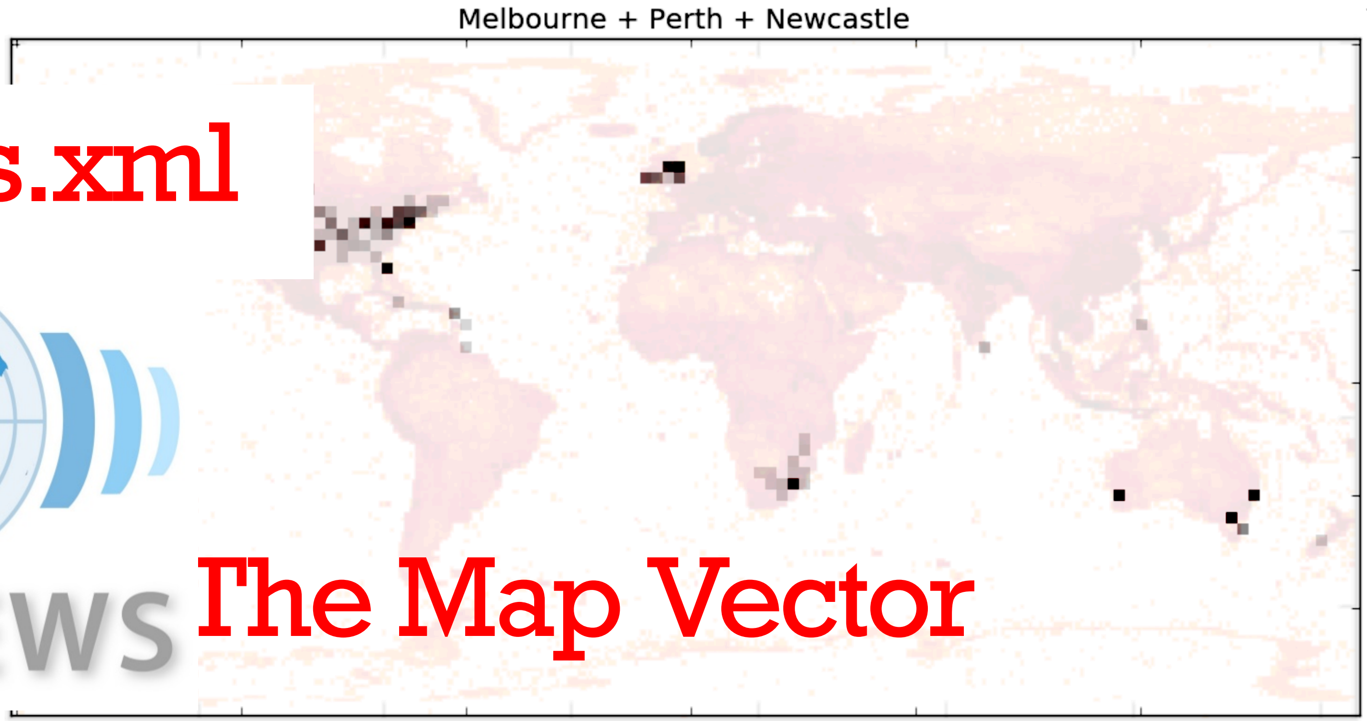
Geocoder	System configuration			Dataset			Average
	Language Features	+	MapVec Features	LGL	WIK	GEO	
CamCoder	CNN		MLP	**0.22**	**0.33**	**0.31**	**0.29**
Lexical Only	CNN		–	0.23	0.39	0.33	0.32
MapVec Only	–		MLP	0.25	0.41	0.32	0.33
Context2vec†	LSTM		MLP	0.24	0.38	0.33	0.32
Context2vec	LSTM		–	0.27	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

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

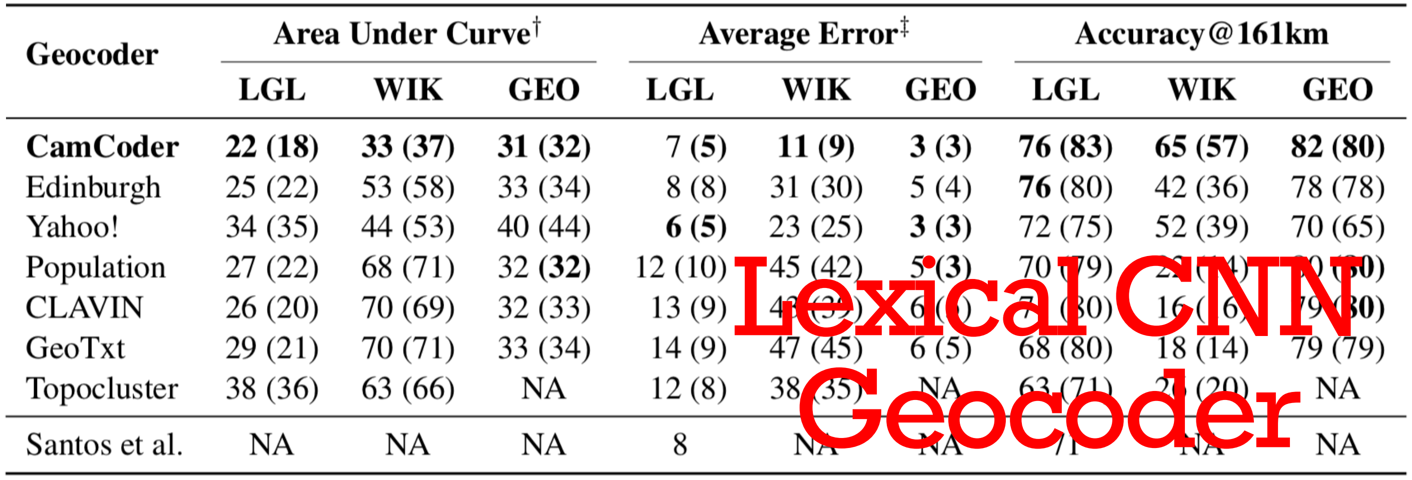
GeoVirus.xml


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

The Map Vector



CamCoder



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 (6)	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

Lexical CNN Geocoder

Money enables much Scientific Research.

Thank You!

www.DREAM-CDT.ac.uk

DREAM

DATA, RISK & ENVIRONMENTAL ANALYTICAL METHODS

50

NERC SCIENCE OF THE ENVIRONMENT

www.NERC.ac.uk

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL

https://ESRC.ukri.org/

Language Technology Lab, University of Cambridge

THANK YOU and CHECK OUT THE PAPER.

https://github.com/milangritta

Questions?

UNIVERSITY OF CAMBRIDGE