LINKING LOCATIONS WITH GEO DATA

Joanne Yang
Joel Nothman, James R. Curran
ə-lab, School of Information Technologies
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGIES

Introducing Locations and Linking

Place names are ambiguous –
synonymy: *Queensland* and *Sunshine State*
multiple referents: *Cambridge, Massachusetts, USA* or *Cambridge, England, UK*.

- Smith and Mann (2003) analysed over 1 million entries in The Getty Thesaurus
- 33% of places in Asia had multiple names
- Almost 30% in Oceania had multiple referents

Named entity linking (NEL) resolves the names of people, places and organisations to a knowledge base (KB) like Wikipedia (Figure 1).

Toponym resolution (TR) focuses on linking place names to their referent (Leidner, 2007). Many systems use gazetteers and machine learning using context for TR and geolocation.

- Better linking of locations may improve NEL
- Motivation: NEL has not exploited **meta-data** used in TR or **location-specific strategies**
- Current state-of-art linking gets about 80%

Analysis of Data and Errors

- Motivated by error analysis of locations in linker
- Almost **40% entities are locations** in TAC 11 data
- Locations account for **almost 50% of errors** – disproportionately difficult

Type	Count	%
Correctly linked named entities	1955	87.1
Locations in TAC 11 data	896	**39.9**
Correctly linked LOCs (% correct)	726	38.7
Location errors (out of all errors)	137	**47.2**

Table 1: *Analysis of* TAC *11 location errors overall*

GeoNames is a geographical database used in TR and in our method:

- Over nine million place names including aliases of cities, countries, mountains, rivers etc
- Each place has administrative level (e.g. country, or level 1 to 4 to represent state)
- Other information like population, capital city, continent included if available
- Users can edit database using web interface
- **Coverage is uneven** because combines multiple knowledge bases, sources

Type	Total #	% zeroes
First order admin (state)	8	0.00
Second order admin	559	2.50
Populated place (PP)	12522	81.30
Section of PP	1977	95.35

Table 2: *GeoNames population counts by administrative code for Australia - uneven by type, higher administrative regions have fewer zeroes*

Disambiguating Location Candidates

Intuition: more populous places are more likely to occur in news text and nearby places likely to occur together. Features:

1. **Highest Population** – link most populous candidate
2. **Hierarchy filtering** – keep candidates in the same country/continent as selected pivot candidate, see Figure 1; experimented with different selections
   e.g. most frequent, highest ranked (top) concept or first concept in document

Hypothesis: improving **location disambiguation and linking impacts NEL** , including non-geographical entities.

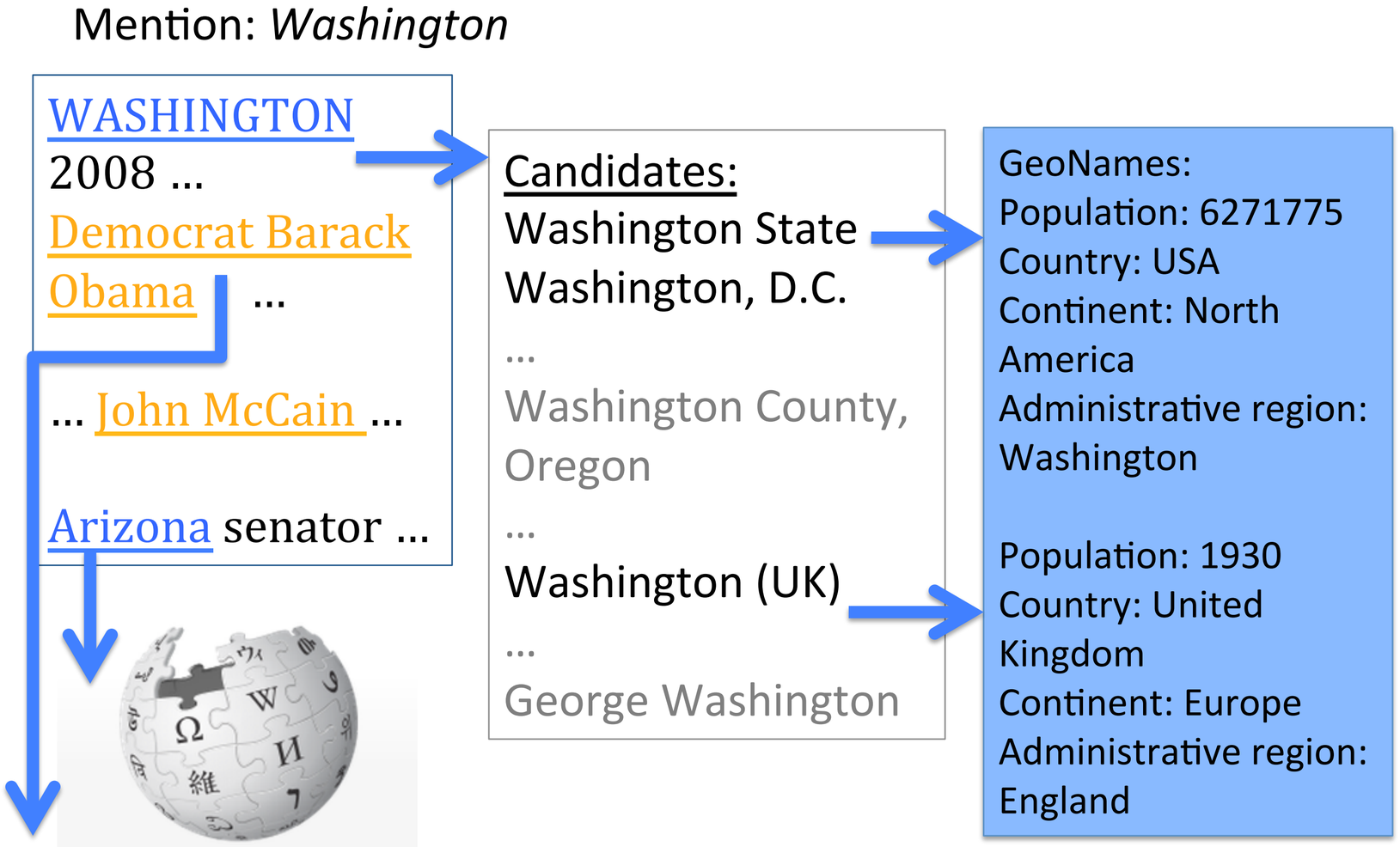

Figure 1: NEL *candidates for* Washington *and how GeoNames can filter by population and hierarchy*

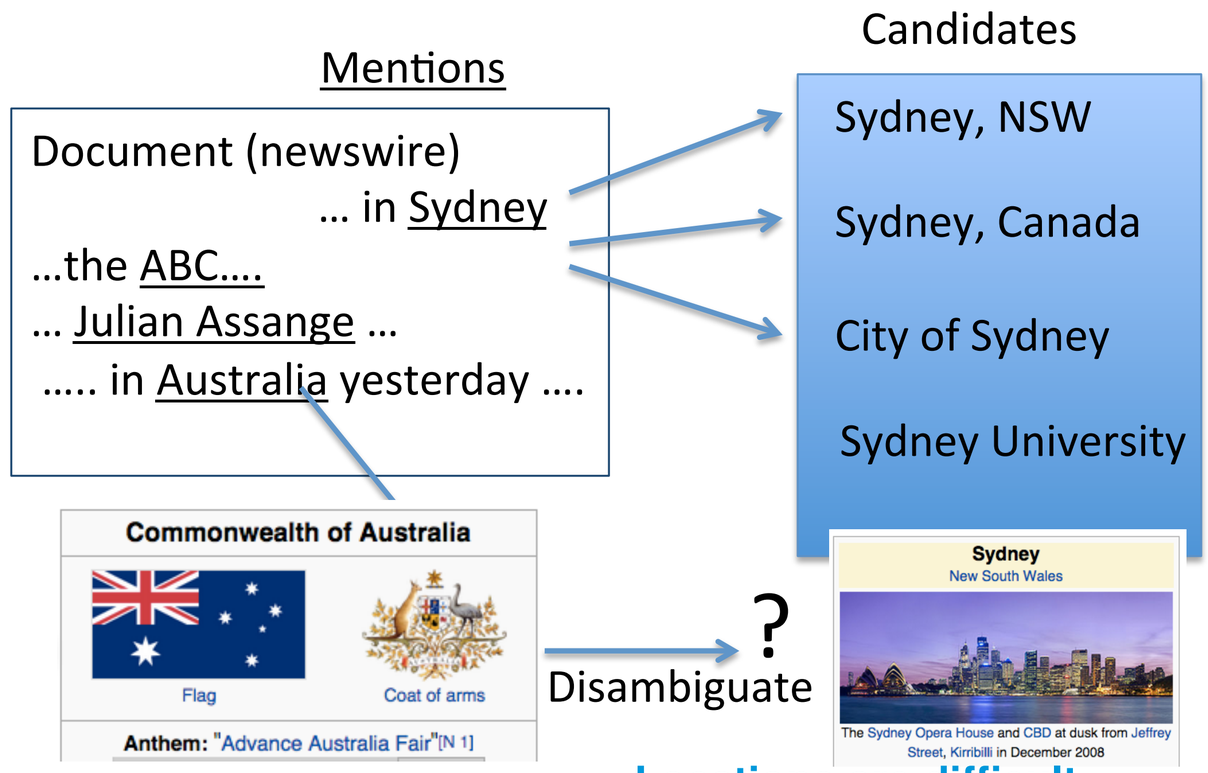

Figure 2: NEL *for Sydney – ambiguity*

Experimental Results

Dataset: TAC 12	Accuracy	Correct LOC
Baseline	74.35	53.20
Highest Population	68.37	25.62
Top Concept in chain	**74.89**	52.95
First Concept Country	**74.57**	52.45

Table 3: TAC *12 results*

Dataset: CN test	Accuracy	Correct LOC
Baseline	70.60	77.15
Top Concept in chain	**71.02**	76.89

Table 4: *CompNews test set results*

- Population counts alone don't work as well
- Country generally better than continent filtering
- **Top concept filtering** performed the best (0.7% improvement on local news training data)

Locations are difficult

Error analysis showed:

- **Missing population**: e.g. *Africa*, small towns
- Population or most frequent incorrect: e.g. *New York State* linked when actually *New York City*
- Local news biased towards smaller locations (American news or CompNews)
- Better filtering for regional? e.g. *Springfield*: 73 entries in GeoNames in US, 236 worldwide

Finding most salient place not easy - headlines and first paragraph can be misleading e.g. He told her he was a young Marine, recently back from **Iraq** is in US.

Future Work

- Use scaled, relative population or combine with other features in supervised linker
- Filter multiple concepts or LOC intersection
- Other methods to determine most salient location or don't filter some e.g. when multiple locations or stories in the same article