NSF IBSS: Location Ontology (San Diego)

Jean Mark Gawron
Dhiraj Patil

San Diego State University

2015 Feb 2

Overview

1. The problem
   - Problem structure
   - Language variation

Location info for non-GPS Tweets

Tweets are associated with a variety of location information besides GPS info, which is present on about only 1% of the Tweets. Location language is one of the most important of these. Such signals are particularly likely to be in important in Tweets concerning spatially and temporally defined events such as disasters and elections.

Topicality of Location

Tweets collected over the period of time during which some extended event is happening will also identify **topical locations**, locations that are important in the development of the event. For example, in a wildfire scenario, fire locations, evacuation areas, or shelters/evacuation centers can become topical locations. Complex spatial language can specify locations with no name. In many cases, the locations referred to in Tweets are not the location of the Tweet sender at the time of sending, so this is a signal that cannot be obtained from GPS Tweets.

Wildfire Tweet

@SDSheriff has issued evacuation calls in the Wintergardens area due to wildfire burning near Aurora Drive & Interstate 8 in Lakeside. (retweeted hundreds of times).

Beyond named locations

The richness of language

2842 West Cortez
Five miles west of rt 125 on rt 52
in the Von's parking lot
near I5

Google geocoding handles this
but not this
or this
nor vagueness, nor extended polygon locations like this

Link Language and space

1. Ontology: For each San Diego Location term
   1. map coordinates or shapefiles
   2. place ontology (the type of place)
   3. subregion relations
2. Language Variation: Two data sources
   1. SD Forum posts
   2. San Diego Wildfire Tweets
3. Future work
   1. Language richness (complex expressions)
   2. Vagueness (richer ontology allowing Gaussian mixtures as locations)

Ontology

Personnel

Jean Mark Gawron	Linguistics
Dhiraj Patil	Computer Science

Resources

Site	Description
`www.census.gov`	Census site gazeteer, national place name and subregion info
`geonames.usgs.gov`	National and state level files for features of various kinds, including populated places, latlons and subregion info
`wiki/XXX,_San_Diego`	Wikipedia pages containing Name, Lat-Long, and Subregion info for many SD neighborhoods
`www.sangis.org`	San Diego specific GIS site, San Diego place names with shapefiles

Ontology entries are **LocationFrames**

BalboaPark	
Semantic Type	Park, Neighborhood
Language	Balboa Park
Lat Long	32.7314 N, 117.1453 W
Subregion of	CentralSanDiego

CentralSanDiego	
Semantic Type	Area
Language	Central San Diego
Lat Long	32.7314 N, 117.1453 W
Subregion of	SanDiego

Further Types

Neighborhoods	Bay Ho, Bay Park, Carmel Valley, Clairemont, Del Mar Heights, Del Mar Mesa, La Jolla, La Jolla Village, Mission Beach, North City, Pacific Beach, Pacific Highlands Ranch, Torrey Hills, Torrey Pines, University City, La Jolla, ...
Parks	Balboa Park, Belmont Park, Black Mountain Open Space Park, Cabrillo National Monument, Chicano Park, Children's Pool Beach, Cowles Mountain, Heritage Park, La Jolla Cove, La Jolla Shores, Los Peasquitos Canyon Preserve, Marine Street Beach, Mission Bay, ...
Hospitals	Alvarado Hospital Medical Center, Scripps Mercy Hospital - Chula Vista, Sharp Coronado Hospital And Healthcare Center, Sharp Mary Birch Hospital For Women, ...

Paths

Paths in San Diego County

NorthSouthFwy	I-5, I-15, I-805, SR 15, SR 67, SR 94, SR 125, SR 163
EastWestFwy	I-8, SR 52, SR 54, SR 56, SR 78,SR 905
Bridges	Cabrillo, Harbor Drive Pedestrian, Lake Hodges, Lilac Road, Los Peasquitos Creek Arch, Pine Valley Creek, San Diego-Coronado
Ports of Entry	Otay Mesa, San Ysidro, Tecate

Problem: Our typical geonames KB sources do NOT have entries for paths. We show later that this matters.

Ontology building: Tasks done

1. Crawled `http://city-data.com/forum/san-diego` and scraped all forum posts
2. Web-scraped `http://en.wikipedia.org/wiki/XXX,_San_Diego` for San Diego area, neighborhood, and district names, subregion relations, and Lat Lons
3. Have merged the Wikipedia and USGS info to one a set of "canonical" place names associated with either lat-long or a subregion relation `http://www.sangis.org/`.

`http://city-data.com`	Scraped .9GB of forum data (19326 files), 7768 place names extracted by the NER (not all SD place names). Built frequency distribution
`/wiki/XXX,_San_Diego`	Scraped 118 SD-related pages, 116 with lat lon info, for a list of canonical SD neighborhood and district names
`geonames.usgs.gov`	Downloaded gazeteer for US populated places (119,115 US places), 416 in SD county, mostly with LatLon.
`www.sangis.org`	Extracted names linked to shapefiles, lat lons or containing regions for Freeway features (2038 freeway features), Hospitals(22), MajorRoads (28,580 road features) Parks (258 Parks), Ports of Entry (6), and miscellaneous places (28,580)
merged	Merged Wikipedia info and USGS info (489 located names)

Place Examples

Oak Ridge Business Center	33.15341949 -117.22333314
Alvarado Hospital Medical Center	6655 ALVARADO ROAD 92120-5298
Black Mtn Nbhd Park	Black Mountain Ranch
Northern SD Neighborhoods	Bay Ho,Bay Park,Carmel Valley,Clairemont,Del Mar Heights,Del Mar Mesa,La Jolla,La Jolla Village,Mission Beach,North City,Pacific Beach,Pacific Highlands Ranch,Torrey Hills,Torrey Pines,University City,Village of La Jolla

Twitter Evaluation

Wildfire data: Pilot study

1. 6500 Tweets collected with the keyword *evacuation* from the 2014 San Diego Wildfire
2. Procedure
   1. Use prefix Trie of location names in location ontology to identify Tweets with substrings which **begin** potential location names, and to activate a set of potential names to match loosely
   2. Run regular expression match on those Tweets against activated names [some trivial variation accounted for: case, omitted spaces]
3. Evaluated robustness of location extraction on Wildfire data by hand annotating Tweets with location names.

Precision/recall (Counting types only)

Compressed names: Size of set of names found in Tweets goes from 805 to 234 if case and hashtag character and spaces are ignored.

	Precision	Recall
compressed	38.7	10.25

Weaknesses of pilot study string matching

KB lacks pathnames	Find pathnames data sources
Ontogy doesn't have **addresses**	Construct address regular expression patterns
Suffix omission **University Boulevard** $\rightarrow$ **University**	Type specific Suffix omission patterns
Robust Abbreviation rules	Abbreviation rules (High School $\rightarrow$ HS, Boulevard $\rightarrow$ Blvd, Torry Pines $\rightarrow$ Torrey Pnes, Willow Grove Elementary School $\rightarrow$ Willow Grove Elem.)
Socially salient locs not in DB	Legoland, Vons, Sears

Precision issues

Precision error: A location reference linked to the wrong KB location or a non location reference misidentified as a location: Many of these errors trace back to poor coverage.

Issue	Example
X + suffix misidentified as X	*Palomar Airport Rd* misidentified as *Palomar Airport* (because we don't know *Palomar Airport Rd*)
complex expression containing X misidentified as X	west of Citracado Parkway
non San Diego names not correctly identified	Bear Valley

Recall issues

Random	Coverage (*Carslbad Sears*, no corresponding entry in our KB). Irregular variant (*tp, tphs*)
Productive	Spelling error. Complex expression (*a half mile west of 5 on Sea World Drive*), *the city of Carlsbad*)

Findings

This pilot study of Tweet location extraction uncovered some of the practical issues in linking real location knowledge bases to actual data using location names. They include

1. Lack of data sources for path names
2. Need for suffix rules allowing optional suffixes (Avenue, Boulevard, etc)
3. Need for abbreviation rules
4. Need for rules for complex location expressions (e.g., "near Aurora Drive & Interstate 8 in Lakeside", "S. of Poinsettia Ln.")
5. Need for data collection to increase set of salient locations (Legoland)
6. Need for address rules

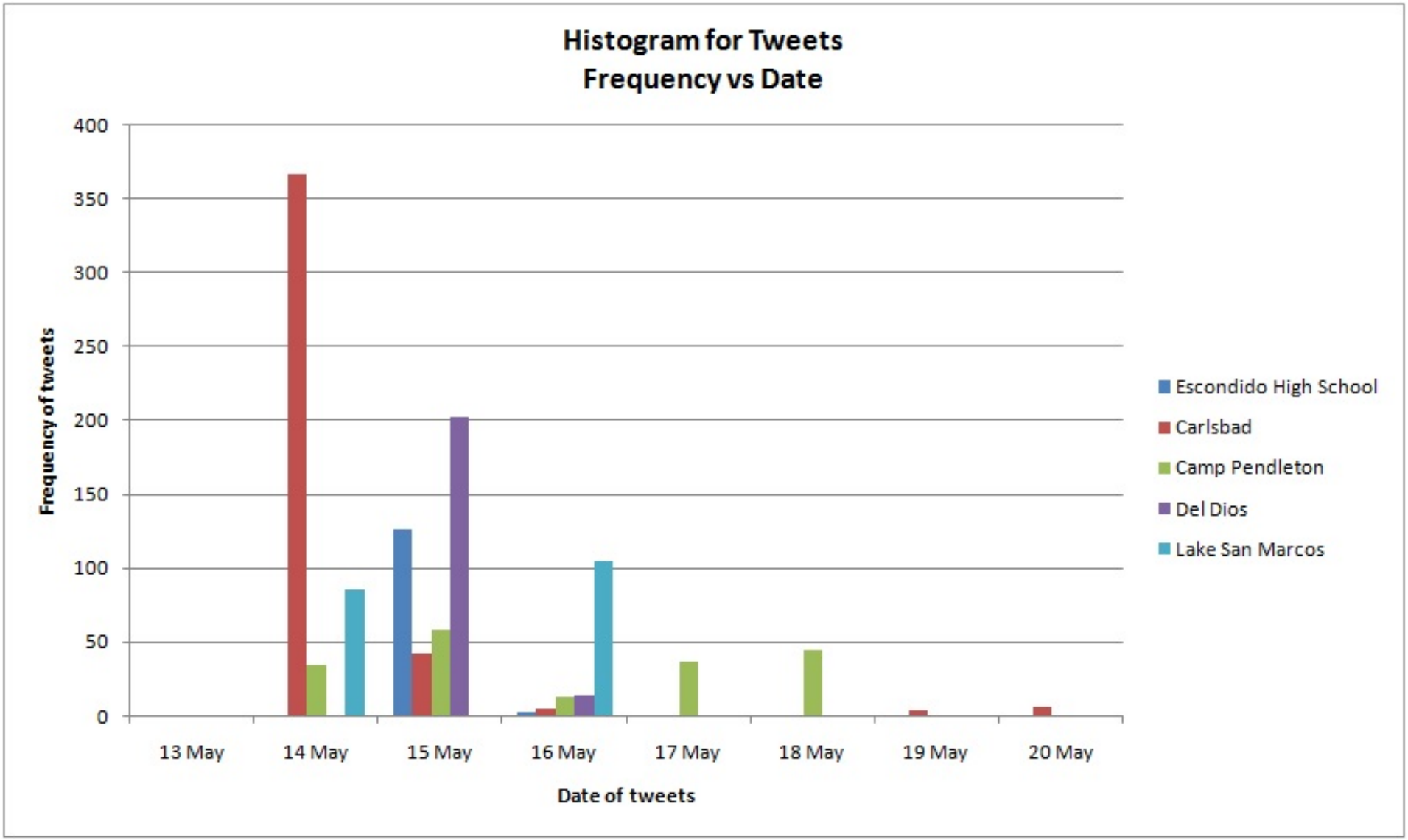
Location Extractor: Version 0.1

1. Implemented suffix/prefix omission rules (optional "Boulevard" in "El Cajon Boulevard", etc.)
2. Extracted path names from the merged geonames KB, wich includes many buildings, and shopping centers, with addresses (*Alvarado Road* from address of *Alvarado Hospital Medical Center*). Merged with a list of San Diego streets from www.geographic.org
3. Implemented address matching: Regular expression matching digits and directions before pathnames (match "2836 West Cortez Street" given "Cortez Street" as a pathname).

Results

	Precision	Recall
compressed	43.4	14.10

Location activity histogram



Left to do

1. Large set of complex expressions
2. Productive abbreviation rules (apply bidirectionally: often the KB entry is abbreviated)
3. Non productive variation: How can we "discover" that 'tphs' is an abbreviation for Torrey Pines High School?
4. Spelling correction

To do: Location Extractor: Version 1.0

The grammar of location expressions

A semantic grammar of location expressions which handles suffixes, prefixes, omissibility, and complex expressions uniformly, creating a semantic specification of a location names, directions, and distances:

$$[\![\text{ west of I5 }]\!] = \mathcal{L} \mid \text{west of}(\mathcal{L}, \text{I5}, [\text{miles } 3])$$