Grounding Text

Jason Baldridge
@jasonbaldridge

Associate Professor
Co-founder & Chief Scientist
What does “barbecue” mean?
What does “barbecue” mean?
What does “barbecue” mean?
What does “barbecue” mean?
What does “barbecue” mean?

Barbecue

[Image: A diagram showing the word 'barbecue' with related terms and a graph of barbecue events over time.]
What does “barbecue” mean?

Barbecue'

[Image: Words associated with barbecue]

[Graph: Number of barbecue events (in Billions)]

© 2013 Jason M Baldridge
What does “barbecue” mean?
What I thought semantics was before 2005

Updated perspective *a la* Ray Mooney (UT Austin CS)

Robocup Simulator

http://www.cs.utexas.edu/users/ml/slides/chen-icml08.ppt
Travel at the Turn of the 20th Century

http://www.lib.utexas.edu/books/travel/index.html
Motivation: Google Lit Trips [http://www.googlelittrips.com/]

Grapes of Wrath in Google Earth

Motivation: Google Lit Trips [http://www.googlelittrips.com/]

Grapes of Wrath in Google Earth

Crisis response: Haiti earthquake

ID: 4025
TITILE: Water needed for woman and child on Delmas 72
DATE: 2010-01-28 22:38:00
LOCATION: Rte Delmas 72
DESCRIPTION:

- English: Need help. I have a 1 year old child and having problems finding drinking water and food. Address is route de delmas 272 Note: this seems like a high number, it might be number 27 of route delmas – Category: 3a. Water shortage
- Creole: Jai besoin daide.jai un enfant de 1 an.il nest pas facile de trouver leau potable et de la nourriture.aidez moi je vous prie.ladresse est: 272,rte de delma
- Comments: In the French, the person talks about the exact address “272 route de delma”; but she doesn’t mention the city which is very tricky for us, thanks
- Ushahidi ID: IDUshahidi: 1986

LATITUDE: 18.5192
LONGITUDE: -72.2831
Crisis response: Haiti earthquake

ID: 4025
TITLE: Water needed for life
DATE: 2010-01-28 22:38
LOCATION: Rte Delmas
DESCRIPTION:

- English: Need help with water and food. This might be number 27.
- Creole: Jai besoin de la nourriture et de l'eau. C'est le numéro 27.
- Comments: In the Red Cross, but she doesn’t know.
- Ushahidi ID: IDUs39

CATEGORY: 2. Urgency
LATITUDE: 18.5192
LONGITUDE: -72.2831
Look, Mom, no hands! (Err, um... no metadata.)

<table>
<thead>
<tr>
<th>WATER</th>
<th>LAND</th>
<th>WRITING</th>
<th>ART</th>
<th>EGYPT</th>
<th>CITY</th>
<th>MECCA</th>
<th>PLAINS</th>
</tr>
</thead>
<tbody>
<tr>
<td>sea</td>
<td>mountain</td>
<td>cloth</td>
<td>church</td>
<td>egypt</td>
<td>street</td>
<td>mecca</td>
<td>city</td>
</tr>
<tr>
<td>coast</td>
<td>feet</td>
<td>book</td>
<td>renaissance</td>
<td>nile</td>
<td>london</td>
<td>city</td>
<td>sioux</td>
</tr>
<tr>
<td>island</td>
<td>rock</td>
<td>illustrations</td>
<td>madonna</td>
<td>cairo</td>
<td>station</td>
<td>place</td>
<td>chicago</td>
</tr>
<tr>
<td>steamer</td>
<td>valley</td>
<td>story</td>
<td>picture</td>
<td>egyptian</td>
<td>railway</td>
<td>day</td>
<td>river</td>
</tr>
<tr>
<td>bay</td>
<td>great</td>
<td>books</td>
<td>venice</td>
<td>temple</td>
<td>road</td>
<td>time</td>
<td>kansas</td>
</tr>
<tr>
<td>land</td>
<td>mountains</td>
<td>work</td>
<td>mark</td>
<td>ancient</td>
<td>hotel</td>
<td>desert</td>
<td>missouri</td>
</tr>
<tr>
<td>ship</td>
<td>hills</td>
<td>life</td>
<td>doge</td>
<td>time</td>
<td>city</td>
<td>mahomet</td>
<td>west</td>
</tr>
<tr>
<td>harbour</td>
<td>rocks</td>
<td>history</td>
<td>early</td>
<td>rameses</td>
<td>square</td>
<td>country</td>
<td>business</td>
</tr>
<tr>
<td>port</td>
<td>side</td>
<td>author</td>
<td>saints</td>
<td>water</td>
<td>place</td>
<td>caravan</td>
<td>miles</td>
</tr>
<tr>
<td>shore</td>
<td>miles</td>
<td>crown</td>
<td>century</td>
<td>desert</td>
<td>house</td>
<td>arabia</td>
<td>hundred</td>
</tr>
<tr>
<td>boat</td>
<td>plain</td>
<td>illustrated</td>
<td>great</td>
<td>tombs</td>
<td>great</td>
<td>journey</td>
<td>line</td>
</tr>
<tr>
<td>water</td>
<td>river</td>
<td>gilt</td>
<td>work</td>
<td>years</td>
<td>english</td>
<td>damascus</td>
<td>dakota</td>
</tr>
<tr>
<td>board</td>
<td>hill</td>
<td>edition</td>
<td>wall</td>
<td>river</td>
<td>office</td>
<td>days</td>
<td>streets</td>
</tr>
<tr>
<td>islands</td>
<td>stone</td>
<td>vols</td>
<td>works</td>
<td>temples</td>
<td>regent</td>
<td>pilgrims</td>
<td>railroad</td>
</tr>
<tr>
<td>town</td>
<td>long</td>
<td>volumee</td>
<td>room</td>
<td>luxor</td>
<td>club</td>
<td>camels</td>
<td>lincoln</td>
</tr>
</tbody>
</table>
Look, Mom, no hands! (Err, um... no metadata.)

<table>
<thead>
<tr>
<th>WATER</th>
<th>LAND</th>
<th>WRITING</th>
<th>ART</th>
<th>EGYPT</th>
<th>CITY</th>
<th>MECCA</th>
<th>PLAINS</th>
</tr>
</thead>
<tbody>
<tr>
<td>sea</td>
<td>mountain</td>
<td>cloth</td>
<td>church</td>
<td>egypt</td>
<td>street</td>
<td>mecca</td>
<td>city</td>
</tr>
<tr>
<td>coast</td>
<td>feet</td>
<td>book</td>
<td>renaissance</td>
<td>nile</td>
<td>london</td>
<td>city</td>
<td>sioux</td>
</tr>
<tr>
<td>island</td>
<td>rock</td>
<td>illustrations</td>
<td>madonna</td>
<td>cairo</td>
<td>station</td>
<td>place</td>
<td>sioux</td>
</tr>
<tr>
<td>steamer</td>
<td>valley</td>
<td>story</td>
<td>picture</td>
<td>egyptian</td>
<td>railway</td>
<td>day</td>
<td>chicago</td>
</tr>
<tr>
<td>bay</td>
<td>great</td>
<td>books</td>
<td>venice</td>
<td>temple</td>
<td>road</td>
<td>time</td>
<td>river</td>
</tr>
<tr>
<td>land</td>
<td>mountains</td>
<td>work</td>
<td>mark</td>
<td>ancient</td>
<td>hotel</td>
<td>desert</td>
<td>kansas</td>
</tr>
<tr>
<td>ship</td>
<td>hills</td>
<td>life</td>
<td>doge</td>
<td>time</td>
<td>city</td>
<td>mahomet</td>
<td>missouri</td>
</tr>
<tr>
<td>harbour</td>
<td>rocks</td>
<td>history</td>
<td>early</td>
<td>rameses</td>
<td>square</td>
<td>country</td>
<td>business</td>
</tr>
<tr>
<td>port</td>
<td>side</td>
<td>author</td>
<td>saints</td>
<td>water</td>
<td>place</td>
<td>caravan</td>
<td>miles</td>
</tr>
<tr>
<td>shore</td>
<td>miles</td>
<td>crown</td>
<td>century</td>
<td>desert</td>
<td>house</td>
<td>arabia</td>
<td>hundred</td>
</tr>
<tr>
<td>boat</td>
<td>plain</td>
<td>illustrated</td>
<td>great</td>
<td>tombs</td>
<td>great</td>
<td>journey</td>
<td>dakota</td>
</tr>
<tr>
<td>water</td>
<td>river</td>
<td>gilt</td>
<td>work</td>
<td>years</td>
<td>english</td>
<td>damascus</td>
<td>dakota</td>
</tr>
<tr>
<td>board</td>
<td>hill</td>
<td>edition</td>
<td>work</td>
<td>river</td>
<td>office</td>
<td>days</td>
<td>streets</td>
</tr>
<tr>
<td>islands</td>
<td>stone</td>
<td>vols</td>
<td>wall</td>
<td>regent</td>
<td>pilgrims</td>
<td>railroad</td>
<td></td>
</tr>
<tr>
<td>town</td>
<td>long</td>
<td>volumee</td>
<td>works</td>
<td>temples</td>
<td>club</td>
<td>camels</td>
<td></td>
</tr>
</tbody>
</table>

Topics with a clear, circumscribed geographic focus **emerge**!
But, of course, metadata is now plentiful.
Geotagged Wikipedia

University of Texas at Austin

From Wikipedia, the free encyclopedia

"University of Texas" redirects here. For the university system, see University of Texas System.

The University of Texas at Austin (University of Texas, UT Austin, UT, or simply Texas) is a public research-intensive university located in Austin, Texas, United States. It is the flagship institution of the University of Texas System.[7][9][10] Founded in 1883, its campus is approximately 0.25 miles (400 m) from the Texas State Capitol in Austin. The institution has the fifth-largest single-campus enrollment in the nation, with over 50,000 undergraduate and graduate students and over 24,000 faculty and staff.[11][12][13][14]

The University of Texas at Austin was named one of the original eight Public Ivy institutions[15] and was inducted into the American Association of Universities in 1925.[16] The university is a major center for academic research, with research expenditures exceeding $640 million for the 2009–2010 school year.[17] The university houses seven museums and seventeen libraries, including the Lyndon Baines Johnson Library and Museum and the Blanton Museum of Art,[18] and operates various auxiliary research facilities, such as the J. J. Pickle Research Campus and the McDonald Observatory. Among university faculty are recipients of the Nobel Prize, Pulitzer Prize, the Wolf Prize, and the National Medal of Science, as well as many other awards.[19]

UT Austin student athletes compete as the Texas Longhorns and are members of the Big 12 Conference. The university has won four NCAA Division I National Football Championships, and has claimed more titles in men's and women's sports than any other school in the Big 12 since the league was founded in 1996.[20] Current and former UT Austin athletes have won 117 Olympic medals, including 14 in Beijing in 2008.[21] The university was recognized by Sports Illustrated as "America's Best Sports College" in 2002.[22]

30° 17′ N   97° 44′ W
Geotagged Twitter

01:55:55  RT @USER_dc5e5498: Drop and give me 50....

05:09:29  I said u got a swisher from redmond!? He said nah kirkland!
         Lol..oooooooooOkay!

05:57:35  Lmao!:) havin a good ol time after work! Unexpected! #goodtimes

06:00:09  RT @USER_d5d93fec: #letsbereal .. No seriously, #letsbereal>>lol.
         Don't start.

06:00:37  On my way to get @USER_60939380 yeee! She want some of this
         strawberry! Sexy!

...

47°3 1’41” N  122°11’52” W
Document geolocation: where is this person?

Philip Flip Kromer @mflip
I'm at Opal Divine's Freehouse (Austin, TX) w/ 2 others
4sq.com/yKFbTB

Philip Flip Kromer @mflip
Heading out w/ @bill_michels and @twbell to Data Drinks at Opal Divines 700 W 6th bit.ly/datadrinks12 - No free booze, just smart geeks

Philip Flip Kromer @mflip
Big data meetup - Hilton 6th floor (@ hilton austin 6th floor meeting rm A) 4sq.com/xxYoIz

Philip Flip Kromer @mflip
@Hipployna @baratunde not to worry, we’re concentrated on making humans smarter - Human + machine way beats machine: collisiondetection.net/mt/archives/20...

Damon Clinkscales @damon
Secret #sxsw tip for coffee lovers.... @HoundstoothATX is downtown on sixth street this week. Thank you @SoundCloud! You chose very well.

Philip Flip Kromer @mflip
@kellm 6:30 Opal Divine's data drinks? /cc @antgoldbloom @marshallk @mattb @caseorganic @i2pi
Language modeling approach

Amsterdam, Zaandam, Amstelveen, Diemen, Landsmeer ...
Frankfurt, Frechen, Hürth, Brühl, Wesseling, ...

Wing & Baldridge 2011: Simple supervised document geolocation with geodesic grids.
Language modeling approach

Amsterdam, Zaandam, Amstelveen, Diemen, Landsmeer ...
Frankfurt, Frechen, Hürth, Brühl, Wesseling, ...

Wing & Baldridge 2011: Simple supervised document geolocation with geodesic grids.
Where’s a word on Earth?

- beach
- mountain
- wine
- barbecue
Where’s a word on Earth?

- beach
- mountain
- wine
- barbecue
Locations of Twitter users are not uniformly distributed!

(Small) GeoUT (Twitter) plotted on Google Earth, one pin per user.

Density of (all) documents in GeoUT over the USA (390 million tweets)
k-d tree for geotagged Wikipedia, looking at N. America

k-d tree for geotagged Wikipedia, looking at N. America

Pre-grid clustering [Erik Skiles, MA thesis, UT Austin, Ling]
Four clusters on GeoUT (390 million tweets)
Four clusters on GeoUT (390 million tweets)

- All tweets
- West coast
- East coast
- Midwest & South
- Spanish language
Automatic document geolocation
Automatic document geolocation

[Serdyukov, Murdock, & van Zwol 2009; Cheng, Caverlee, & Lee 2010; Wing & Baldridge 2011]
Image geo-location: http://graphics.cs.cmu.edu/projects/im2gps/
### Performance (kd-tree with clustering)

<table>
<thead>
<tr>
<th>Wikipedia (entire world)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Half of documents geotagged within <strong>12 km</strong> of truth</td>
</tr>
<tr>
<td>Percent of documents within 166km (100 miles): <strong>91%</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Twitter (USA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Half of users geotagged within <strong>330 km</strong> of truth</td>
</tr>
<tr>
<td>Percent of documents within 166km (100 miles): <strong>40%</strong></td>
</tr>
</tbody>
</table>

For better or worse, it soon might not matter whether you have location turned on or not... what you say is where you are / are from. (Also, other factors, e.g. who you are linked to, of course.)
Hierarchical geo-location with logistic regression

Wing & Baldridge 2014: Hierarchical Discriminative Classification for Text-Based Geolocation.
Performance (kd-tree with clustering)

**Twitter (USA)**
Half of users geotagged within **170 km** of truth
Percent of documents within 166km (100 miles): **49%**

**Twitter (World)**
Half of users geotagged within **490 km** of truth
Percent of documents within 166km (100 miles): **31%**

**Flickr (entire world)**
Half of documents geotagged within **18 km** of truth
Percent of documents within 166km (100 miles): **66%**
### Hierarchical logistic regression beats flat naive Bayes

#### Accuracy @ 161 km, kd-tree grid

<table>
<thead>
<tr>
<th></th>
<th>Naive Bayes</th>
<th>Hierarchical LR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Twitter USA</strong></td>
<td>36.2</td>
<td>49.2</td>
</tr>
<tr>
<td><strong>Twitter World</strong></td>
<td>28.7</td>
<td>31.3</td>
</tr>
<tr>
<td><strong>Flickr</strong></td>
<td>58.5</td>
<td>66.0</td>
</tr>
<tr>
<td><strong>English Wikipedia</strong></td>
<td>84.5</td>
<td>88.9</td>
</tr>
<tr>
<td><strong>German Wikipedia</strong></td>
<td>89.3</td>
<td>90.2</td>
</tr>
<tr>
<td><strong>Portuguese Wikipedia</strong></td>
<td>77.1</td>
<td>89.5</td>
</tr>
</tbody>
</table>
Logistic regression weights good features heavily
They visit Portland every year.
Toponym (place name) resolution

They visit **Portland** every year.

Which Portland? (Also: Canada, Australia, Ireland...)

© 2013 Jason M Baldridge
Toponym resolution in context

Although Elisha Newman made the first land entry in the township of Portland (June, 1833), he did not become a settler until three years later, by which time a few settlers had located in the town. From Mr. Newman's story, it appears that early in 1833, he was visiting friends in Ann Arbor, and during an evening conversation discussed with others the subject of unlocated lands lying west of Ann Arbor. One of the company (Joseph Wood) remarked that he had been out with the party sent to survey Ionia and other counties, and that the surveyors were struck by the valuable water-power at the mouth of the Looking Glass River, saying there would surely be a village there some day.

Mr. Newman was at once taken with the idea of locating lands at the mouth of the Looking Glass. Following up his impulse, he made ready to start at once, and, accompanied by James Newman and Joseph Wood, went out to the Looking Glass on a tour of inspection. Being satisfied with the location, he returned Eastward with his companions, and at White Pigeon made his land entry.

Newman did not return for a permanent settlement until the spring of 1836, and meanwhile, in November, 1833, Philo Bogue bought a piece of land on section 28, in the bend of the Grand River, where he proposed to set up a trading post. Unaided he rolled up a log cabin near where the Detroit, Lansing, and Northern depot was located, and when he brought the house into decent shape went over to Hunt's at Lyons for his family, whom he had left there against such time as he should have affairs prepared for their comfort.
Although Elisha Newman made the first land entry in the township of Portland (June, 1833), he did not become a settler until three years later, by which time a few settlers had located in the town. From Mr. Newman's story, it appears that early in 1833, he was visiting friends in Ann Arbor, and during an evening conversation discussed with others the subject of unlocated lands lying west of Ann Arbor. One of the company (Joseph Wood) remarked that he had been out with the party sent to survey Ionia and other counties, and that the surveyors were struck by the valuable water-power at the mouth of the Looking Glass River, saying there would surely be a village there some day.

Mr. Newman was at once taken with the idea of locating lands at the mouth of the Looking Glass. Following up his impulse, he made ready to start at once, and, accompanied by James Newman and Joseph Wood, went out to the Looking Glass on a tour of inspection. Being satisfied with the location, he returned Eastward with his companions, and at White Pigeon made his land entry.

Newman did not return for a permanent settlement until the spring of 1836, and meanwhile, in November, 1833, Philo Bogue bought a piece of land on section 28, in the bend of the Grand River, where he proposed to set up a trading post. Unaided he rolled up a log cabin near where the Detroit, Lansing, and Northern depot was located, and when he brought the house into decent shape went over to Hunt's at Lyons for his family, whom he had left there against such time as he should have affairs prepared for their comfort.
Although Elisha Newman made the first land entry in the township of Portland (June, 1833), he did not become a settler until three years later, by which time a few settlers had located in the town. From Mr. Newman's story, it appears that early in 1833, he was visiting friends in Ann Arbor, and during an evening conversation discussed with others the subject of unlocated lands lying west of Ann Arbor. One of the company (Joseph Wood) remarked that he had been out with the party sent to survey Ionia and other counties, and that the surveyors were struck by the valuable water-power at the mouth of the Looking Glass River, saying there would surely be a village there some day.

Mr. Newman was at once taken with the idea of locating lands at the mouth of the Looking Glass. Following up his impulse, he made ready to start at once, and, accompanied by James Newman and Joseph Wood, went out to the Looking Glass on a tour of inspection. Being satisfied with the location, he returned Eastward with his companions, and at White Pigeon made his land entry.

Newman did not return for a permanent settlement until the spring of 1836, and meanwhile, in November, 1833, Philo Bogue bought a piece of land on section 28, in the bend of the Grand River, where he proposed to set up a trading post. Unaided he rolled up a log cabin near where the Detroit, Lansing, and Northern depot was located, and when he brought the house into decent shape went over to Hunt's at Lyons for his family, whom he had left there against such time as he should have affairs prepared for their comfort.
Spatial minimality

Although Elisha Newman made the first land entry in the township of Portland (June, 1833), he did not become a settler until three years later, by which time a few settlers had located in the town. From Mr. Newman’s story, it appears that early in 1833, he was visiting friends in Ann Arbor, and during an evening conversation discussed with others the subject of unlocated lands lying west of Ann Arbor. One of the company (Joseph Wood) remarked that he had been out with the party sent to survey Ionia and other counties, and that the surveyors were struck by the valuable water-power at the mouth of the Looking Glass River, saying there would surely be a village there some day.

Mr. Newman was at once taken with the idea of locating lands at the mouth of the Looking Glass. Following up his impulse, he made ready to start at once, and, accompanied by James Newman and Joseph Wood, went out to the Looking Glass on a tour of inspection. Being satisfied with the location, he returned Eastward with his companions, and at White Pigeon made his land entry.

Newman did not return for a permanent settlement until the spring of 1836, and meanwhile, in November, 1833, Philo Bogue bought a piece of land on section 28, in the bend of the Grand River, where he proposed to set up a trading post. Unaided he rolled up a log cabin near where the Detroit, Lansing, and Northern depot was located, and when he brought the house into decent shape went over to Hunt’s at Lyons for his family, whom he had left there against such time as he should have affairs prepared for their comfort.
Although Elisha Newman made the first land entry in the township of Portland (June, 1833), he did not become a settler until three years later, by which time a few settlers had located in the town. From Mr. Newman's story, it appears that early in 1833, he was visiting friends in Ann Arbor, and during an evening conversation discussed with others the subject of unlocated lands lying west of Ann Arbor. One of the company (Joseph Wood) remarked that he had been out with the party sent to survey Ionia and other counties, and that the surveyors were struck by the valuable water-power at the mouth of the Looking Glass River, saying there would surely be a village there some day.

Mr. Newman was at once taken with the idea of locating lands at the mouth of the Looking Glass. Following up his impulse, he made ready to start at once, and, accompanied by James Newman and Joseph Wood, went out to the Looking Glass on a tour of inspection. Being satisfied with the location, he returned Eastward with his companions, and at White Pigeon made his land entry.

Newman did not return for a permanent settlement until the spring of 1836, and meanwhile, in November, 1833, Philo Bogue bought a piece of land on section 28, in the bend of the Grand River, where he proposed to set up a trading post. Unaided he rolled up a log cabin near where the Detroit, Lansing, and Northern depot was located, and when he brought the house into decent shape went over to Hunt's at Lyons for his family, whom he had left there against such time as he should have affairs prepared for their comfort.
Spatial minimality often fails

I moved from Encinitas, CA, a nice beach town in North San Diego County to Asheville, NC. By far, Ashville is more hip, especially West Asheville. Asheville has a lot in common with Portland. Austin, I've never been to so I cannot comment. But what makes a place cool and hip, in my opinion are that give a area "punch". There are a lot of ingredients. One is geography. Add a college or university (and all that they bring- and draw), good restaurants, a good music scene, a progressive attitude and tolerance. Hmmm. I'm sure there are many more to ponder. But that's my start. Oh, lots of bars!


City-data.com incorrectly marks “West” and “Portland” as the cities in Texas -- presumably because of their textual and spatial proximity to “Austin”.
Spatial minimality often fails

I moved from Encinitas, CA, a nice beach town in North San Diego County to Asheville, NC. By far, Ashville is more hip, especially West Asheville. Asheville has a lot in common with Portland. Austin, I've never been to so I cannot comment. But what makes a place cool and hip, in my opinion are that give a area "punch". There are a lot of ingredients. One is geography. Add a college or university (and all that they bring- and draw), good restaurants, a good music scene, a progressive attitude and tolerance. Hmmm. I'm sure there are many more to ponder. But that's my start. Oh, lots of bars!


City-data.com incorrectly marks “West” and “Portland” as the cities in Texas -- presumably because of their textual and spatial proximity to “Austin”.

But: it is clear from the text that Portland, Oregon and Austin, Texas are the referents, though their states are never mentioned and are far from the other locations!
Toponym classifiers

Strategy: build a textual classifier per toponym by obtaining *indirectly* labeled examples from Wikipedia.
Toponym classifiers

Strategy: build a textual classifier per toponym by obtaining \textit{indirectly} labeled examples from Wikipedia.

\textbf{Portland Youth Philharmonic}

\textit{From Wikipedia, the free encyclopedia}

The Portland Youth Philharmonic (PYP) is the oldest youth orchestra in the United States, established in 1924 as the Portland Junior Symphony (PJS). Now based in Portland, Oregon, the orchestra’s origin dates back to 1910 when music teacher Mary V. Dodge began playing music for local children in Burns. Dodge purchased instruments for the children and organized the orchestra which would become known as the \textit{Sagebrush}.

\textbf{Widgery Wharf}

\textit{From Wikipedia, the free encyclopedia}

\textit{Widgery Wharf} is an historic wharf in Portland, Maine. Built in in the late 1700s, the wharf is named for the Widgery family which controlled the local molasses trade at the time of completion. Members of the Widgery family include Congressman William Widgery. The precise date of the wharf's building is unknown. With
Toponym classifiers

Strategy: build a textual classifier per toponym by obtaining *indirectly* labeled examples from Wikipedia.

---

**Portland Youth Philharmonic**

From Wikipedia, the free encyclopedia

The Portland Youth Philharmonic (PYP) is the oldest youth orchestra in the United States, established in 1924 as the Portland Junior Symphony (PJS). Now based in Portland, Oregon, the orchestra's origin dates back to 1910 when music teacher Mary V. Dodge began playing music for local children in Burns. Dodge purchased instruments for the children and organized the orchestra which would become known as the Sagebrush.

---

**Widgery Wharf**

From Wikipedia, the free encyclopedia

Widgery Wharf is an historic wharf in Portland, Maine. Built in in the late 1700s, the wharf is named for the Widgery family which controlled the local molasses trade at the time of completion. Members of the Widgery family include Congressman William Widgery. The precise date of the wharf's building is unknown.
Strategy: build a textual classifier per toponym by obtaining *indirectly* labeled examples from Wikipedia.
Toponym classifiers

Strategy: build a textual classifier per toponym by obtaining *indirectly* labeled examples from Wikipedia.
Strategy: build a textual classifier per toponym by obtaining *indirectly* labeled examples from Wikipedia.

**Portland Youth Philharmonic**

The Portland Youth Philharmonic (YP) is the oldest youth orchestra in the United States, established in 1924 as the Portland Junior Symphony (PJS). Now based in Portland, Oregon, the orchestra's origin dates back to 1910 when music teacher Mary V. Dodge began playing music for local children in Burns. Dodge purchased instruments for the children and organized the orchestra which would become known as the Sagebrush.

**Wigdery Wharf**

Wigdery Wharf is an historic wharf in Portland, Maine. Built in the late 1700s, the wharf is named for the Wigdery family which controlled the local molasses trade at the time of completion. Members of the Wigdery family include Congressman William Wigdery. The precise date of the wharf's building is unknown.
Toponym classifiers

Strategy: build a textual classifier per toponym by obtaining indirectly labeled examples from Wikipedia.

\[
P(\text{Portland-OR} \mid \text{music}) > P(\text{Portland-ME} \mid \text{music}) \\
P(\text{Portland-OR} \mid \text{wharf}) < P(\text{Portland-ME} \mid \text{wharf})
\]
# Results: disambiguating toponyms

<table>
<thead>
<tr>
<th>Population</th>
<th>Average error distance</th>
<th>Accuracy</th>
<th>Average error distance</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TR-CoNLL</strong></td>
<td></td>
<td></td>
<td><strong>Perseus Civil War Corpus</strong></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>August 1996</td>
<td>Late 19th Century</td>
</tr>
<tr>
<td>Reuters News Texts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SPIDER</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(spatial minimality)</td>
<td>216</td>
<td>81.0</td>
<td>1749</td>
<td>59.7</td>
</tr>
<tr>
<td><strong>WISTR</strong></td>
<td>2180</td>
<td>30.9</td>
<td>266</td>
<td>57.5</td>
</tr>
<tr>
<td>(Wiki supervised)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>SPIDER + WISTR</strong></td>
<td>279</td>
<td>82.3</td>
<td>855</td>
<td>69.1</td>
</tr>
<tr>
<td></td>
<td>430</td>
<td>81.8</td>
<td><strong>201</strong></td>
<td><strong>85.9</strong></td>
</tr>
</tbody>
</table>

**Take-home message:** text classifiers are very effective & can be boosted by spatial minimality algorithms.
Identifying, disambiguating, and displaying toponyms
Grounding often involves connecting text to knowledge sources and other modalities (image, video) & bootstrapping.
Grounding often involves connecting text to knowledge sources and other modalities (image, video) & bootstrapping.

Also, they can help us create models for deeper aspects of language, such as syntactic structure and logical form.
Lexical brain decoding [Yarkoni, Poldrack, Nichols, Van Essen & Wager (2011)]

Term-based search

"Pain"

Related studies

Mechanisms of Directed

An fMRI Investigation of

Placebo-Induced Changes in fMRI
In the Anticipation and
Experience of Pain

Automated coordinate
eXtraction

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>-23</td>
<td>18</td>
<td>45</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>3</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>-40</td>
<td>0</td>
<td>-16</td>
<td>1</td>
</tr>
<tr>
<td>35</td>
<td>-41</td>
<td>29</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>18</td>
<td>33</td>
<td>2</td>
</tr>
</tbody>
</table>

Meta-analysis

P(Pain|Activation)
Lexical brain decoding [Yarkoni, Poldrack, Nichols, Van Essen & Wager (2011)]
He says, she says  http://www.tweetolife.com/gender/
Temporality of words, by hour http://www.tweetolife.com/hour/
Temporality of words, by hour http://www.tweetolife.com/hour/
Temporality of expressions, by day: http://www.google.com/trends
Temporality of expressions, by day: http://www.google.com/trends
Temporality of expressions, by year: http://ngrams.googlelabs.com/
Temporal resolution [Kumar, Lease, and Baldridge 2011]
Temporal resolution [Kumar, Lease, and Baldridge 2011]
Temporal resolution [Kumar, Lease, and Baldridge 2011]
Temporal resolution [Kumar, Lease, and Baldridge 2011]
Temporal resolution [Kumar, Lease, and Baldridge 2011]
Temporal resolution [Kumar, Lease, and Baldridge 2011]
More modalities: videos [Motwani & Mooney, 2012]

Training Video

* A woman is riding horse in a beach.
* A woman is riding on a horse.
* A woman is riding on a horse.

NL description

STIP features

Discover Activity Label

Classifier Trained on input features as STIP features and classes as activity cluster labels

ride, walk, run, move, race
Beyond word co-occurrences for vector-space models

<table>
<thead>
<tr>
<th></th>
<th>bear</th>
<th>boat</th>
<th>car</th>
<th>cow</th>
<th>hadoop</th>
<th>snow</th>
<th>water</th>
<th>wrench</th>
</tr>
</thead>
<tbody>
<tr>
<td>beach</td>
<td>3</td>
<td>234</td>
<td>42</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>325</td>
<td>0</td>
</tr>
</tbody>
</table>
### Beyond word co-occurrences for vector-space models

<table>
<thead>
<tr>
<th></th>
<th>bear</th>
<th>boat</th>
<th>car</th>
<th>cow</th>
<th>hadoop</th>
<th>snow</th>
<th>water</th>
<th>wrench</th>
</tr>
</thead>
<tbody>
<tr>
<td>beach</td>
<td>3</td>
<td>234</td>
<td>42</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>325</td>
<td>0</td>
</tr>
</tbody>
</table>

---

**Text Analytics Summit, June 2013**

© 2013 Jason M Baldridge
# Beyond word co-occurrences for vector-space models

<table>
<thead>
<tr>
<th></th>
<th>bear</th>
<th>boat</th>
<th>car</th>
<th>cow</th>
<th>hadoop</th>
<th>snow</th>
<th>water</th>
<th>wrench</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>234</td>
<td>42</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>325</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Beyond word co-occurences for vector-space models

<table>
<thead>
<tr>
<th></th>
<th>bear</th>
<th>boat</th>
<th>car</th>
<th>cow</th>
<th>hadoop</th>
<th>snow</th>
<th>water</th>
<th>wrench</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>234</td>
<td>42</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>325</td>
<td>0</td>
</tr>
</tbody>
</table>

beach

![Map of a location with markers indicating various points of interest.](image_url)
Beyond word co-occurences for vector-space models

<table>
<thead>
<tr>
<th></th>
<th>bear</th>
<th>boat</th>
<th>car</th>
<th>cow</th>
<th>hadoop</th>
<th>snow</th>
<th>water</th>
<th>wrench</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>234</td>
<td>42</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>325</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>
Beyond word co-occurrences for vector-space models

<table>
<thead>
<tr>
<th>bear</th>
<th>boat</th>
<th>car</th>
<th>cow</th>
<th>hadoop</th>
<th>snow</th>
<th>water</th>
<th>wrench</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>234</td>
<td>42</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>325</td>
<td>0</td>
</tr>
</tbody>
</table>

beach
Beyond word co-occurences for vector-space models

<table>
<thead>
<tr>
<th>bear</th>
<th>boat</th>
<th>car</th>
<th>cow</th>
<th>hadoop</th>
<th>snow</th>
<th>water</th>
<th>wrench</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>234</td>
<td>42</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>325</td>
<td>0</td>
</tr>
</tbody>
</table>

![Map](image1.png)

![Graph](image2.png)

![Graph](image3.png)
Beyond word co-occurrence for vector-space models

<table>
<thead>
<tr>
<th>bear</th>
<th>boat</th>
<th>car</th>
<th>cow</th>
<th>hadoop</th>
<th>snow</th>
<th>water</th>
<th>wrench</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>234</td>
<td>42</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>325</td>
<td>0</td>
</tr>
</tbody>
</table>

beach

© 2013 Jason M Baldridge
Beyond word co-occurrences for vector-space models

<table>
<thead>
<tr>
<th></th>
<th>bear</th>
<th>boat</th>
<th>car</th>
<th>cow</th>
<th>hadoop</th>
<th>snow</th>
<th>water</th>
<th>wrench</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>3</td>
<td>234</td>
<td>42</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>325</td>
<td>0</td>
</tr>
</tbody>
</table>

beach

© 2013 Jason M Baldridge

Text Analytics Summit, June 2013
Combining distributional models with logics

Erk (2013): “Towards a semantics for distributional representations.”

Garrette et al (2012): “A formal approach to linking logical form and vector-space lexical semantics”
Multi-component structured vector-space models

the children visit the beach

children → visit → beach

Agent

Patient
Language learning in context [Kim & Mooney, 2013]

"at the very next intersection take a right onto the plain path and follow it to the end of the hall"
Language learning in context [Kim & Mooney, 2013]

"at the very next intersection take a right onto the plain path and follow it to the end of the hall"

Travel ( steps: 1 ),
Verify ( side: CONCRETE HALLWAY ),
Turn ( RIGHT ),
Verify ( back: WALL, front: CONCRETE HALLWAY
left: EASEL ),
Travel ( steps: 1 ),
Verify ( front: WALL )
All your meaning are belong to us

"software understands the meaning"
All your meaning are belong to us

"software understands the meaning"
All your meaning are belong to us

"software understands the meaning"
Grounding matters

http://davidrothman.net/2009/09/02/all-your-healthbase-are-belong-to-us-want-em-back/
Open Source Software (Scala/Java)

ScalaNLP

- Junto - label propagation
  https://github.com/scalanlp/junto

- Chalk - NLP
  https://github.com/scalanlp/chalk

- Nak - machine learning
  https://github.com/scalanlp/nak

- Breeze - linear algebra
  https://github.com/scalanlp/nak

OpenNLP - machine learning / NLP
http://opennlp.apache.org/

- Textgrounder - document geolocation
  https://github.com/utcompling/textgrounder

- Fieldspring - toponym resolution
  https://github.com/utcompling/fieldspring

- Low-resource POS tagging
  https://github.com/dhgarrette/low-resource-pos-tagging-2013

- Updown - polarity classification
  https://github.com/scalanlp/junto
Open Source Software (Scala/Java)

ScalaNLP

- OpenNLP - machine learning / NLP
  http://opennlp.apache.org/
- Junto - label propagation
  https://github.com/scalanlp/junto
- Chalk - NLP
  https://github.com/scalanlp/chalk
- Nak - machine learning
  https://github.com/scalanlp/nak
- Breeze - linear algebra
  https://github.com/scalanlp/nak
- Textgrounder - document geolocation
  https://github.com/utcompling/textgrounder
- Fieldspring - toponym resolution
  https://github.com/utcompling/fieldspring
- Low-resource POS tagging
  https://github.com/dhgarrette/low-resource-pos-tagging-2013
- Updown - polarity classification
  https://github.com/scalanlp/junto

Friday, September 5, 14
This research was sponsored by:

Grant from the Morris Memorial Trust Fund

Grant: W911NF-10-1-0533

Final note: Whitman had it right many years ago!

A SONG of the rolling earth, and of words according,
Were you thinking that those were the words, those upright lines? those curves, angles, dots?
No, those are not the words, the substantial words are in the ground and sea,
They are in the air, they are in you.

- Walt Whitman, A Song of the Rolling Earth (in Leaves of Grass)
Supervision
- documents labeled with latitude & longitude

Methods
- Language Modeling for Information Retrieval

Code
- Textgrounder: https://github.com/utcompling/textgrounderer

Publications
Supervision
- indirectly acquired toponym annotations using a gazetteer and geo-annotated Wikipedia

Methods
- logistic regression
- named entity recognition

Code
- Fieldspring: https://github.com/utcompling/fieldspring

Publications