

Multifaceted Toponym Recognition for Streaming News

Michael D. Lieberman Hanan Samet

Center for Automation Research, Institute for Advanced Computer Studies,
Department of Computer Science, University of Maryland,
College Park, MD 20742 USA
{codepoet,hjs}@cs.umd.edu

July 27, 2011

Streaming News

- Explosion of digitization: Lots of data!
 - News constantly being created in a 24-hour news cycle
 - Continuous publishing model
 - Non-traditional news sources: bloggers, Twitter
 - Web-capable mobile devices can access and generate news
- Collectively can be considered as a constant stream of news to be processed and understood, to enable its spatio-textual retrieval
- Challenges:
 - Staying up-to-date with latest data
 - Traditional database designs not intended to deal with rapidly changing datasets
 - Coordinating a complex process of news processing
 - Enabling fast spatial retrieval of large amounts of news data
 - Performance evaluations involving streaming news
 - Corpora: Usually have only a few articles from one or two prominent news sources (e.g., NY Times)
 - Not representative of Internet news which by far consists of smaller, local news sources

Geography in Text

- News often has a strong geographic component which is useful for geographic retrieval of news
- Spatial data is specified using text (called *toponyms*) rather than geometry, which means that there is some ambiguity involved
- Advantage: From a geometric standpoint, the textual specification captures both the point and spatial extent interpretations of the data
 - City can be specified by either a point such as its centroid, or a region corresponding to its boundary, depending on zoom level
- One disadvantage: We are not always sure if a term is a geographic location or not (e.g., does “Jordan” refer to a country or is it a surname as in “Michael Jordan”?)
- Another disadvantage: If a geographic location, then which, if any, of the possibly many instances of geographic locations with the same name is meant (e.g., does “London” refer to an instance in the UK, Ontario, Canada, or one of many others?)

Geotagging

- Must understand the geographic content of each article
- Geotagging: Convert textual specifications of geographic locations found in free running text into their lat/long representations
 - E.g., “Paris, France” → “48.87, 2.36”
- Geotagging a text document consists of:
 1. *Toponym recognition*: Finding all textual references to geographic locations (*toponyms*)
 2. *Toponym resolution*: Choosing the correct location interpretation (i.e., lat/long values) for each toponym
- Core challenge: Resolving ambiguities in textual location specifications
 - E.g., “Paris”: “Paris, France”, “Paris, Texas”, or “Paris Hilton”?
- Geotagging enables unambiguous spatial indexing and retrieval of text documents using locations present in the text
 - More informative than simply using user’s or news source’s location, if present
 - Requires deeper understanding of document’s content

Multifaceted Toponym Recognition

- Use evidence from a wide variety of sources to capture as many potential toponyms as possible
- Leverage the strengths of several different approaches
 - I.e., rule-based and machine learning-based methods
- Generally heuristic in nature

- Main concern: high toponym recall
 - I.e., missing as few toponyms in documents as possible
 - Toponym precision is restored by later geotagging process

- Primary contributions:
 - Comprehensive multifaceted toponym recognition method designed for streaming news that uses many types of evidence
 - Novel experimental evaluation of our methods, using corpora of streaming news, and compared against two prominent competitors

Talk Outline

1. NewsStand system
2. Finding toponyms
3. Filtering out toponyms
4. Evaluation on streaming news

Talk Outline

1. NewsStand system
2. Finding toponyms
3. Filtering out toponyms
4. Evaluation on streaming news

NewsStand

- Toponym recognition methods employed in our system named NewsStand [Teitler et al., 2008]
- Enables people to search for news using a map query interface
- Advantage: A map, coupled with an ability to vary the zoom level at which it is viewed, provides an inherent granularity to the search process that facilitates an approximate spatial search
- Distinguished from today's prevalent keyword-based conventional search methods that provide a very limited facility for approximate spatial searches
 - Realized by permitting a match via use of a subset of keywords
 - Users have little grasp of which spatial keywords to use
- Map query interface requires no spatial keywords
 - Act of pointing at a location and selecting zoom level permits approximate spatial search without the use of keywords

B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. NewsStand: A new view on news. In *GIS'08: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 144–153, Irvine, CA, November 2008.

Live Demo

NewsStand is available at <http://newsstand.umiacs.umd.edu>

NewsStand Summary

1. Crawls the web looking for news sources and feeds
 - Indexing 8,000 news sources
 - About 50,000 news articles per day
2. Aggregate news articles by both content similarity and location
 - Articles about the same event are grouped into clusters
3. Rank clusters by importance which is based on:
 - Number of articles in cluster
 - Number of unique newspapers in cluster
 - Event's rate of propagation to other newspapers
4. Associate each cluster with its geographic focus or foci
5. Display each cluster at the positions of the geographic foci
6. Other options:
 - (a) Topic type (e.g., General, Business, Sports, Entertainment)
 - (b) Image and video galleries
 - (c) Map stories by people, disease. . .
 - (d) User-generated news (e.g., social networks such as Twitter)

Talk Outline

1. NewsStand system
2. Finding toponyms
3. Filtering out toponyms
4. Evaluation on streaming news

Running Example

- Excerpt from an article in the Paris News about a local politician campaigning in Paris, Texas
- Mentions multiple places in Texas

Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

Running Example

- Excerpt from an article in the Paris News about a local politician campaigning in Paris, Texas
- Mentions multiple places in Texas

*Democratic candidate for **Texas** Railroad Commissioner Jeff Weems stumped in **Paris** late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-**Paris**, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A **Houston** attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled **Lamar County** “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in **Austin**. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of **Dish, Texas**, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of **Paris**.*

- True toponyms: **Texas, Paris, Houston, Lamar County, Austin, Dish**

Running Example

- Excerpt from an article in the Paris News about a local politician campaigning in Paris, Texas
- Mentions multiple places in Texas

*Democratic candidate for **Texas** Railroad Commissioner Jeff **Weems** stumped in **Paris** late **Friday** in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark **Homer**, D-**Paris**, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A **Houston** attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — **Weems** labeled **Lamar County** “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in **Austin**. Although he did not accuse incumbents of wrong doing, **Weems** said he is upset about the handling of a complaint by the mayor of **Dish, Texas**, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of **Paris**.*

- True toponyms: **Texas, Paris, Houston, Lamar County, Austin, Dish**
- Potential mistakes: **Weems, Homer, Friday** (all in Texas)

Finding Toponyms

- Create initial, large set of potential toponyms using a variety of methods and techniques
- Rule-based methods:
 1. Entity dictionary matching
 2. Cue word matching
 3. Toponym refactoring
- Machine learning-based methods:
 1. Off-the-shelf NER software with postprocessing
 2. Part of speech tagging
- Goal: Maximize toponym recall
 - Recognition precision will be restored in later stages of processing

Entity Tables

- Find entities in document from curated lists of entities corresponding to locations, and other types
 - Knowledge of non-location entities can inform toponym recognition
 - E.g., “Apple” as a company, rather than a small city in Ohio
- Also find instances of *cue words* that signal entities of various types
 - E.g., “County of X”
- Entities selected from problematic entities in NewsStand
- These lists can never be complete, but they serve as a useful starting point for toponym recognition

General entities		Spatial cues	
Religion	Christian, Islam, Hindu	Populated regions	State of X
Season	Spring, Fall	Populated places	Town of X, Y City
Direction	South, Northeast, Midwest	Comma groups	X and Y counties
Day	Monday, Friday	Water features	Gulf of X, Y Lake
Month	March, August	Spot features	X School, Mt. Y
Timezone	EST, WEST	Universities	University of X at Y
Color	Gray, Navy, Lime	General	X-based, Y-area
Organization entities		Person entities	
Brand names	Apple, Coke, Toyota	Honorifics	Mr. X; Ms Y; Dr. Z
News agencies	AP, UPI	Generational suffixes	X, Jr.; Y III
Terror groups	Hamas, Taliban	Postnominals	X, KBE; Y, M.D.
Unions	NEA, PETA	Job titles	Sen. X; President Y; Sgt. Z
Government orgs	Congress, Army	Declaratory words	X said; added Y
Postnominals	X Corp., Y Inc.	Common given names	John X; Jennifer Y

Statistical Tools

- Incorporate NLP tools that train and use statistical language models (HMMs, CRFs)
- Use:
 1. Part of speech (POS) tagging
 - Location names tend to be proper nouns
 - Assign grammatical part of speech to each input token
 - Collect groups of proper nouns as toponyms, which results in high recall (will miss few toponyms) and low precision (many reported toponyms will be wrong)
 - Use TreeTagger [Schmid, 1994] trained on Penn TreeBank
 2. Named entity recognition (NER)
 - Generalization of toponym recognition to arbitrary entities
 - Collect reported location entities as toponyms
 - Associate scores with entities
 - Use Stanford NER [Finkel et al., 2005] with default model trained on CoNLL, MUC-5, MUC-7, and ACE data, yielding person, location, and organization entities

Postprocessing Filters

- Use postprocessing steps to enable us to avoid common pitfalls with which statistical NER tools have trouble
- Attempt to correct entity boundaries that caused the NER system to incorrectly fragment entities
 - E.g., “Equatorial [LOC Guinea]” vs. “[LOC Equatorial Guinea]”
- Solution: Find other instances in document corresponding to entire entity phrase, and expand entity boundaries if found
- Articles often mention the same entity multiple times but only fully specify the entity the first time, which causes the NER system to commit type errors
 - E.g., [PER Paul Washington] vs. [LOC Washington]
- Solution: Correct type errors for fragments of these entities found elsewhere in document by finding matching prefixes and suffixes

Running Example

Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

1. Initial text

Running Example

Democratic candidate for **Texas** Railroad Commissioner Jeff Weems stumped in Paris late **Friday** in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, **Texas**, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

1. Initial text

2. **Entity tables**: [LOC **Texas**], [PER Jeff Weems], [DAY **Friday**], [PER Mark Homer]

Running Example

Democratic candidate for **Texas** Railroad Commissioner **Jeff Weems** stumped in Paris late **Friday** in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state **Rep. Mark Homer**, **D-Paris**, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled **Lamar County** “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, **Texas**, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

1. Initial text
2. **Entity tables**: [LOC **Texas**], [PER **Jeff Weems**], [DAY **Friday**], [PER **Mark Homer**]
3. **Cue words**: Rep. [PER **Mark Homer**], D-[LOC **Paris**], [LOC **Lamar County**]

Running Example

Democratic candidate for **Texas Railroad Commissioner Jeff Weems** stumped in **Paris** late **Friday** in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state **Rep. Mark Homer, D-Paris**, said it will be refreshing to have someone on the **Railroad Commission** who “has a concept of what those people are there for.” A **Houston** attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — **Weems** labeled **Lamar County** “ground zero” for **Democrats** winning statewide elections before telling his audience what he plans to do differently in **Austin**. Although he did not accuse incumbents of wrong doing, **Weems** said he is upset about the handling of a complaint by the mayor of **Dish, Texas**, the site of a gas compressor station. That station is similar to the **Midcontinent Express Pipeline** compressor station south of **Paris**.

1. Initial text
2. **Entity tables**: [LOC Texas], [PER Jeff Weems], [DAY Friday], [PER Mark Homer]
3. **Cue words**: Rep. [PER Mark Homer], D-[LOC Paris], [LOC Lamar County]
4. **Proper noun phrases**: [NP Democratic], [NP Railroad Commissioner Jeff Weems], [NP Paris], [NP Rep. Mark Homer], [NP Railroad Commission], [NP Houston], [NP Weems], [NP Lamar County], [NP Democrats], [NP Austin], [NP Dish], [NP Texas], [NP Midcontinent Express Pipeline]

Running Example

Democratic candidate for *Texas Railroad Commissioner Jeff Weems* stumped in *Paris* late *Friday* in the *Precinct 5, Place 1 Justice of the Peace* courtroom where he spoke to about 25 people. In introductory remarks, state *Rep. Mark Homer, D-Paris*, said it will be refreshing to have someone on the *Railroad Commission* who “has a concept of what those people are there for.” A *Houston* attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — *Weems* labeled *Lamar County* “ground zero” for *Democrats* winning statewide elections before telling his audience what he plans to do differently in *Austin*. Although he did not accuse incumbents of wrong doing, *Weems* said he is upset about the handling of a complaint by the mayor of *Dish, Texas*, the site of a gas compressor station. That station is similar to the *Midcontinent Express Pipeline* compressor station south of *Paris*.

1. Initial text

2. **Entity tables:** [LOC Texas], [PER Jeff Weems], [DAY Friday], [PER Mark Homer]

3. **Cue words:** Rep. [PER Mark Homer], D-[LOC Paris], [LOC Lamar County]

4. **Proper noun phrases:** [NP Democratic], [NP Railroad Commissioner Jeff Weems], [NP Paris], [NP Rep. Mark Homer], [NP Railroad Commission], [NP Houston], [NP Weems], [NP Lamar County], [NP Democrats], [NP Austin], [NP Dish], [NP Texas], [NP Midcontinent Express Pipeline]

5. **Named-entity recognition:**

[PER Jeff Weems]	0.999	[LOC Houston]	0.917
[LOC Paris]	0.997	[PER Weems]	0.849
[ORG Railroad Commission]	0.995	[LOC Lamar County]	0.737
[LOC Austin]	0.995	[LOC Texas]	0.557
[ORG Midcontinent Express Pipeline]	0.973	[ORG Democratic]	0.539
[PER Mark Homer]	0.920		

Talk Outline

1. NewsStand system
2. Finding toponyms
3. Filtering out toponyms
4. Evaluation on streaming news

Filtering Toponyms

- After initial recognition phase, toponyms are noisy and contain many non-toponyms
- Here, execute additional logic that remove the most egregious errors, while not overly impacting final toponym recall
- Act as additional postprocessing filters on the entire recognition process

- Methods:
 - Ensure toponym qualifier consistency
 - Active verbs
 - Noun adjuncts
 - Type propagation

Ensuring Toponym Qualifier Consistency (Refactoring)

- Location names can be referred to in multiple ways, and cue word positions can vary by locale
 - E.g., “Prince George’s County” (Maryland) vs. “County Kildare” (Ireland)
 - E.g., “County Kildare” vs. “Co. Kildare”
- Account for these variations by refactoring toponym names via pattern matching to associate additional names with each entity
- Refactoring allows more chances for successful lookup in the gazetteer

First name		Second name
Co. X	→	County X
Dr. X	→	Doctor X
Ft. X	→	Fort X
Mt. X	→	Mount X
St. X	→	Saint X
X Co.	→	X County
X Twp.	→	X Township
X County	↔	County X
X County	↔	County of X
X Lake	↔	Lake X
X Parish	↔	Parish of X
X Township	↔	Township of X
X SchType	→	X SchType School

Active Verbs Imply Non-Locations

- Many non-location entities tend to be *active*, i.e., they perform actions (e.g., people, organizations), while locations tend to be *passive*, i.e., they do not
 - E.g., a person would “say” something, while in general, a location would not
- Use the POS tagger to find grammatical subjects of active voice verbs, and disqualify them as locations
- Search for location entities following active verbs and reset their types to proper noun phrase
 - Does not determine the exact type of entity, but exact type is not necessary since we are interested in locations
- Caveat: Does not account for *metonymy* in toponyms where a location name is used to refer to a non-location entity
 - E.g., “Washington stated on Monday. . .” where “Washington” refers to the US government
 - However, repeated instances of “Washington” would provide evidence of these errors since metonyms are relatively uncommon in text

Noun Adjuncts

- Determining the type of location evidence to use in resolving locations can be difficult
 - E.g., “In Russia, US officials. . .”
 - Both “Russia” and “US” are countries
 - But, can be mistaken for *object/container* evidence (a pair of toponyms, one of which contains the other) and “Russia” may be mistaken for any of several places named “Russia” in the US

- To resolve this ambiguity, give priority to *noun adjunct* evidence (i.e., nouns that function as adjectives by modifying other nearby nouns) over *object/container* evidence
 - E.g., in our example, “US” modifies “officials”
 - Using “US” in *object/container* evidence is not warranted

Type Propagation

- Leverage the “one sense per discourse” assumption [Gale et al., 1992] to group entities together into equivalence classes
 - E.g., all instances of “Washington” grouped together
- If all entities in group g are either untyped or of a consistent type t , set types of all entities in g to t , otherwise do nothing
 - E.g., “Washington”: 2 “PER”, 3 “LOC”: Do nothing
 - E.g., “Washington”: 2 “PER”, 3 untyped: Set all to “PER”
- Limits errors as compared to majority voting scheme

Running Example

Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled Lamar County “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

Running Example

*Democratic candidate for Texas Railroad Commissioner Jeff Weems stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — Weems labeled **Lamar County** “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, Weems said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.*

1. **Toponym refactoring**: [LOC Lamar County] → [LOC County of Lamar]

Running Example

Democratic candidate for Texas Railroad Commissioner **Jeff Weems** stumped in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A Houston attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — **Weems** labeled **Lamar County** “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, **Weems** said he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

1. **Toponym refactoring**: [LOC Lamar County] → [LOC County of Lamar]
2. **Active verbs**: [PER Jeff Weems] stumped, [PER Weems] labeled, [PER Weems] said

Running Example

Democratic candidate for Texas Railroad Commissioner **Jeff Weems stumped** in Paris late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-Paris, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A **Houston attorney** with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — **Weems labeled Lamar County** “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in Austin. Although he did not accuse incumbents of wrong doing, **Weems said** he is upset about the handling of a complaint by the mayor of Dish, Texas, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of Paris.

1. **Toponym refactoring**: [LOC Lamar County] → [LOC County of Lamar]
2. **Active verbs**: [PER Jeff Weems] stumped, [PER Weems] labeled, [PER Weems] said
3. **Noun adjuncts**: [LOC Houston] attorney

Running Example

Democratic candidate for **Texas** Railroad Commissioner **Jeff Weems** **stumped** in **Paris** late Friday in the Precinct 5, Place 1 Justice of the Peace courtroom where he spoke to about 25 people. In introductory remarks, state Rep. Mark Homer, D-**Paris**, said it will be refreshing to have someone on the Railroad Commission who “has a concept of what those people are there for.” A **Houston** attorney with life-long experience in the energy business — first as an oil field worker and now representing both oil and gas firms as well as landowners — **Weems** **labeled** **Lamar County** “ground zero” for Democrats winning statewide elections before telling his audience what he plans to do differently in **Austin**. Although he did not accuse incumbents of wrong doing, **Weems** **said** he is upset about the handling of a complaint by the mayor of **Dish, Texas**, the site of a gas compressor station. That station is similar to the Midcontinent Express Pipeline compressor station south of **Paris**.

1. **Toponym refactoring**: [LOC Lamar County] → [LOC County of Lamar]
2. **Active verbs**: [PER Jeff Weems] stumped, [PER Weems] labeled, [PER Weems] said
3. **Noun adjuncts**: [LOC Houston] attorney
4. **Final location entities**: Texas, Paris, Houston, Lamar County, Austin, Dish

Talk Outline

1. NewsStand system
2. Finding toponyms
3. Filtering out toponyms
4. Evaluation on streaming news

Evaluation

- Implemented our toponym recognition methods in the NewsStand system
- Compared our own methods with two prominent competitors:
 - Thomson Reuters's OpenCalais
 - Yahoo! Placemaker
- These are full geotagging systems (i.e., perform toponym recognition and resolution) but we only use recognition when evaluating their performance
- Neither we nor they provide a means of tuning precision/recall tradeoff
- Performed evaluations on a new corpus of streaming news gathered from NewsStand
 - Contrasts with conventional evaluations that use small, static, homogenous corpora of news

Existing Corpora

- We gathered statistics about existing corpora used in geotagging research
- Relatively small sizes of corpora
 - About 400 documents on average
 - Contrast with the roughly 40k articles retrieved by NewsStand in just a single day
- Homogeneous: Contain articles from only a single news source
 - E.g., Reuters, NY Times
- Average number of toponyms per document (T / D) is fairly consistent (7–8)

Work	Source	Docs	Topos	T / D
Amitay et al. [2004]	Web pages	600	7082	11.8
Buscaldi and Magnini [2010]	L'Adige	150	1042	6.9
Buscaldi and Rosso [2008]	GeoSemCor	186	1210	6.5
Garbin and Mani [2005]	Gigaword	165	1275	7.7
Leidner [2006]	RCV1	946	6980	7.4
Lieberman et al. [2010]	LGL	588	4793	8.2
Manov et al. [2003]	News	101	792	7.8
Overell and Ruger [2008]	Wikipedia	1000	1395	1.4
Roberts et al. [2010]	ACE'05	369	5562	15.1
Volz et al. [2007]	Reuters	250	?	?
	Average	436	3348	8.1

Toponym Statistics

- Track toponym counts and other statistics from streaming news articles processed by NewsStand, over time
- Goal: Determine whether NewsStand's toponym recognition procedure has good day-to-day performance in terms of expected recall
- Procedure:
 - Sample seven days' worth of news from November 2010, limiting samples to articles with at least 300 words
 - Execute toponym recognition on news articles collected on each day
 - Count number of toponyms found on each day
- Can be applied easily and automatically to large collections of articles

■ Evaluation results:

- Majority of sampled days have topos/doc between 7.2–7.5, which falls in our expected range of 7–8
- Weekends (06 Nov, 28 Nov) have different publication pattern
- Large number of articles from a variety of sources, in contrast to existing news corpora

Date	Docs	Sources	Topos	T / D
02 Nov 2010	27591	2086	207110	7.5
06 Nov 2010	13355	1245	124430	9.3
10 Nov 2010	28795	2182	208366	7.2
15 Nov 2010	26052	1952	195669	7.5
19 Nov 2010	24193	2018	173630	7.2
23 Nov 2010	26937	2067	194804	7.2
28 Nov 2010	14245	1250	148996	10.5

Streaming News Corpora

- To evaluate toponym recognition accuracy requires corpora of documents annotated with toponyms
- Used two corpora in our evaluation:
 - *LGL*
 - Introduced in our previous work [Lieberman et al., 2010]
 - Intended to be a collection of news articles from smaller news sources, rather than major sources as in prior work of others
 - 621 articles from 114 local newspapers
 - Useful for testing accuracy on a variety of small news sources
 - *Clust*
 - A new corpus created for this work
 - Intended to capture streaming news about large, major news events often published in multiple sources
- Together, *LGL* and *Clust* allow evaluation on small and large streaming news stories, respectively

Streaming News Corpora

- To create *Clust*, selected clusters with 5–100 articles and had articles from at least four unique news sources, sampled between January and April 2010
 - Ensures reasonable variation in articles over time and across news sources/audiences
 - Contains stories with more journalistic impact
- For each cluster, randomly selected one article for manual annotation

■ Comparison of *LGL* and *Clust*

- Toponyms in *LGL* correspond to smaller places, while those in *Clust* are larger places
- Comparable number of toponyms per article in both corpora
- *Clust* contains roughly twice as many articles as *LGL*

	<i>LGL</i>	<i>Clust</i>
Articles	621	13327
News sources	114	1607
Annotated docs	621	1080
Annotated topos	4765	11564
Distinct topos	1177	2320
Median topos per doc	6	8
Location types:		
Total topos	4765	11564
City	2287	3837
≥ 100k pop	756	2377
< 100k pop	1531	1460
Country	911	3540
State	784	2487
County	525	519

Toponym Accuracy

- Using annotated corpora, determine how well the toponym recognition procedure finds toponyms
- Measure performance in terms of *precision* (of all toponyms reported, how many are correct) and *recall* (of all toponyms, how many were reported)
- Need to account for gazetteer differences and slight differences in recognition methods
 - E.g., ground truth “[_{LOC} New York state]” vs. system-generated “[_{LOC} New York] state”
- Consider two ways of matching ground truth and system-generated toponyms
 - *Exact matching*: Toponym boundaries must coincide
 - *Overlap matching*: Toponyms are allowed to overlap
- System-generated toponym is correct, but considered incorrect using exact matching, and correct using overlap matching
- Also, for each method, consider an additional variant that removes toponyms if not present in our gazetteer
 - Denoted with “_{Gaz}”, e.g., “NewsStand_{Gaz}”, and termed *gazetteer filtering*

Toponym Accuracy: LGL

- NewsStand variants greatly outperform OpenCalais and Placemaker in terms of toponym recall, due to the latter two systems' missing many of the toponyms (small $|S|$)
 - At least 0.10 recall, and in some cases 0.20
- Comes at expense of precision, which can be restored in later stages of processing
 - E.g., with gazetteer filtering (NewsStand_{Gaz}), precision jumps greatly with little decrease in recall
- OpenCalais and Placemaker seemed biased toward toponym precision
- Placemaker is greatly affected by gazetteer filtering, while OpenCalais is not
 - Seems to indicate different gazetteers and matching differences
- Performance of all methods are comparable in terms of F_1 score

	$ S $	$ G \cap S $ E / O	Precision E / O	Recall E / O	F_1 E / O
NewsStand	23345	3879 / 4645	0.166 / 0.199	0.814 / 0.975	0.276 / 0.331
NewsStand _{Gaz}	5960	3619 / 3738	0.607 / 0.627	0.759 / 0.784	0.675 / 0.697
OpenCalais	1959	1830 / 1871	0.934 / 0.955	0.384 / 0.393	0.544 / 0.557
OpenCalais _{Gaz}	1873	1757 / 1791	0.938 / 0.956	0.369 / 0.376	0.530 / 0.540
Placemaker	4593	3129 / 3683	0.681 / 0.802	0.657 / 0.773	0.669 / 0.787
Placemaker _{Gaz}	3796	3013 / 3112	0.794 / 0.820	0.632 / 0.653	0.704 / 0.727

Toponym Accuracy: Clust

- NewsStand variants again outperform OpenCalais and Placemaker by even larger margins, again due to OpenCalais and Placemaker missing many of the toponyms (small $|S|$)
- Performance scores for *Clust* are generally higher than *LGL*
 - Indicates that *Clust*'s toponyms are easier to recognize, likely due to greater presence of common toponyms, e.g., country names

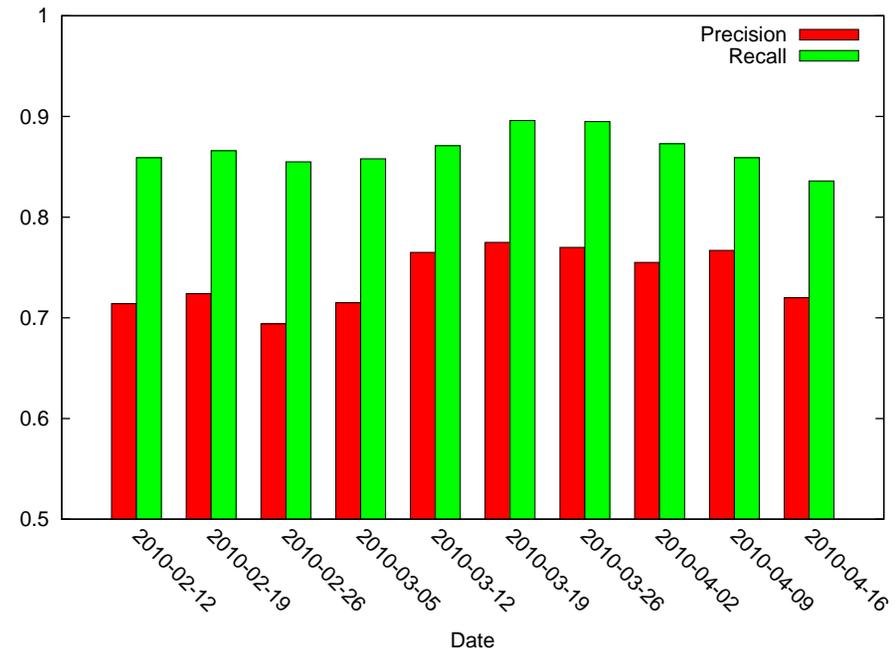
	$ S $	$ G \cap S $ E / O	Precision E / O	Recall E / O	F_1 E / O
NewsStand	44184	10243 / 11330	0.232 / 0.256	0.886 / 0.980	0.368 / 0.406
NewsStand _{Gaz}	13589	9909 / 10036	0.729 / 0.739	0.857 / 0.868	0.788 / 0.798
OpenCalais	6452	6208 / 6326	0.962 / 0.980	0.537 / 0.547	0.689 / 0.702
OpenCalais _{Gaz}	6060	5843 / 5941	0.964 / 0.980	0.505 / 0.514	0.663 / 0.674
Placemaker	9796	6782 / 8549	0.692 / 0.873	0.586 / 0.739	0.635 / 0.800
Placemaker _{Gaz}	7466	6469 / 6593	0.866 / 0.883	0.559 / 0.570	0.679 / 0.693

Streaming Evaluation

- Evaluation on entire static corpus does not reflect day-to-day performance which is characteristic of streaming news
- To measure streaming performance, we split *Clust* into weekly samples and measured precision and recall for NewsStand_{Gaz} using overlap matching

■ Results:

- Performance is relatively consistent over all time periods: mean of 0.739 precision and 0.868 recall
- Indicates that NewsStand's toponym recognition is well suited for streaming news



Future Work

- Leverage NewsStand's clustering module to improve toponym recognition of documents in clusters
 - E.g., “Mr. Washington” in one document in a cluster provides evidence that “Washington” in another document refers to a person, not a location
- Examine usage patterns of individual heuristics used in toponym recognition to determine their frequency of use and overall utility
- Encode our heuristics as features to be used in machine learning techniques, e.g., coreference analysis
 - Feature weights determine usefulness of individual heuristics
 - Adjust weights using NewsStand's error feedback mechanism
- Perform *toponym-centric* evaluation rather than *document-centric* evaluation
 - Select set of highly ambiguous words that can be interpreted as toponyms
 - Annotate only these toponyms, rather than entire documents
 - Evaluate performance on these toponyms in a large set of documents over time
- More suited for streaming news, as toponyms appear in many usage scenarios
- Can annotate a large number of documents quickly
- Use NewsStand's error feedback mechanism to determine words for the corpus and to create annotations

Conclusion

- Toponym recognition is vital to enable geospatial retrieval applications
- Locations are specified using text, rather than geometry, and recognizing toponyms involves resolving ambiguities present in textual specifications of locations
- Multifaceted toponym recognition methods ensure high recall and reasonable precision
- As more news sources move online, algorithms tailored for streaming news will become more important
- Thanks for your attention!
- And to our sponsor:



References

- E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-Where: Geotagging web content. In *SIGIR'04: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 273–280, Sheffield, UK, July 2004.
- D. Buscaldi and B. Magnini. Grounding toponyms in an Italian local news corpus. In *GIR'10: Proceedings of the 6th Workshop on Geographic Information Retrieval*, Zurich, Switzerland, February 2010.
- D. Buscaldi and P. Rosso. A conceptual density-based approach for the disambiguation of toponyms. *IJGIS: International Journal of Geographical Information Science*, 22(3):301–313, March 2008.
- J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In *ACL'05: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 363–370, Ann Arbor, MI, June 2005.
- W. A. Gale, K. W. Church, and D. Yarowsky. One sense per discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, 1992.
- E. Garbin and I. Mani. Disambiguating toponyms in news. In *HLT/EMNLP'05: Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 363–370, Vancouver, Canada, October 2005.
- J. L. Leidner. An evaluation dataset for the toponym resolution task. *CEUS: Computers, Environment, and Urban Systems*, 30(4):400–417, July 2006.
- M. D. Lieberman, H. Samet, and J. Sankaranarayanan. Geotagging with local lexicons to build indexes for textually-specified spatial data. In *ICDE'10: Proceedings of the 26th International Conference on Data Engineering*, pages 201–212, Long Beach, CA, March 2010.
- D. Manov, A. Kiryakov, B. Popov, K. Bontcheva, D. Maynard, and H. Cunningham. Experiments with geographic knowledge for information extraction. In *Proceedings of the HLT-NAACL 2003 Workshop on Analysis of Geographic References*, pages 1–9, Edmonton, Canada, May 2003.
- S. E. Overell and S. Rüger. Using co-occurrence models for placename disambiguation. *IJGIS: International Journal of Geographical Information Science*, 22(3):265–287, March 2008.
- K. Roberts, C. A. Bejan, and S. Harabagiu. Toponym disambiguation using events. In *FLAIRS'10: Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference*, pages 271–276, Daytona Beach, FL, May 2010.
- H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, pages 154–164, Manchester, UK, September 1994.
- B. E. Teitler, M. D. Lieberman, D. Panozzo, J. Sankaranarayanan, H. Samet, and J. Sperling. NewsStand: A new view on news. In *GIS'08: Proceedings of the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 144–153, Irvine, CA, November 2008.
- R. Volz, J. Kleb, and W. Mueller. Towards ontology-based disambiguation of geographical identifiers. In *I3'07: Proceedings of the WWW 2007 Workshop on I3: Identity, Identifiers, Identification, Entity-Centric Approaches to Information and Knowledge Management on the Web*, Banff, Canada, May 2007.