SMASAC - Entity Extraction/Named Entity Recognition (NER)

GRÉGOIRE BUREL, MAYANK KEJRIWAL (PEDRO SZEKELY) AND PRASHANT KHARE

Gather data from various information sources

Extract key information, verify trustworthiness, classify, etc.

Connect and **Visualise** information.

Named Entity Recognition (NER)

- NER is a classic problem in the NLP literature
  - Decades of research, with recent methods including deep learning

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell–Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:
LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE

(Nadeau and Sakine, 2007)

Named Entity Recognition (NER)

- NER is a classic problem in the NLP literature
  - Decades of research, with recent methods including deep learning
- Social media involves unique NER challenges due to irregular text

1	The Hobbit has FINALLY started filming! I cannot wait!
2	Yess! Yess! Its official Nintendo announced today that they Will release the Nintendo 3DS in north America march 27 for $250
3	Government confirms blast n nuclear plants n japan...don't knw wht s gona happen nw...

(Ritter et al., 2011)

Named Entity Recognition (NER)

- NER is a classic problem in the NLP literature
  - Decades of research, with recent methods including deep learning
- Social media involves unique NER challenges due to irregular text
- Crisis data is even more difficult due to presence of ‘uncommon’ entity types (e.g., weather warnings)

(Ritter et al., 2011)

Definition: NER

- Given a set of *entity types* (e.g., PERSON, LOCATION, ORGANIZATION...) and a text corpus, automatically detect and extract typed instances (entities) from the text
  - The finer-grained the types (or the ontology), the harder the problem!

Motivation for NER

- Many named entities in tweets and social media

Natl Hurricane Ctr Retweeted

NHC Atlantic Ops @NHC_Atlantic · 6 Oct 2017

A Storm Surge Warning has been issued from Morgan City Louisiana eastward to the Alabama/Florida border for #Nate. hurricanes.gov

Natl Hurricane Ctr @NWSNHC · 20 Sep 2017

Hurricane #Maria made landfall near Yabucoa, Puerto Rico around 6:15am AST with maximum sustained winds of 155 mph (250 km/h) @NWS @NOAA

Motivation for NER

- Many named entities mentioned in tweets and other social media
- Extracting such entities (and also relations) enables *semantic search* and analytics applications
  - What locations have received ‘Storm Surge Warnings’ from the NHC in the last 10 days?
  - What organizations were involved in relief efforts for Hurricane Irma?

Motivation for NER

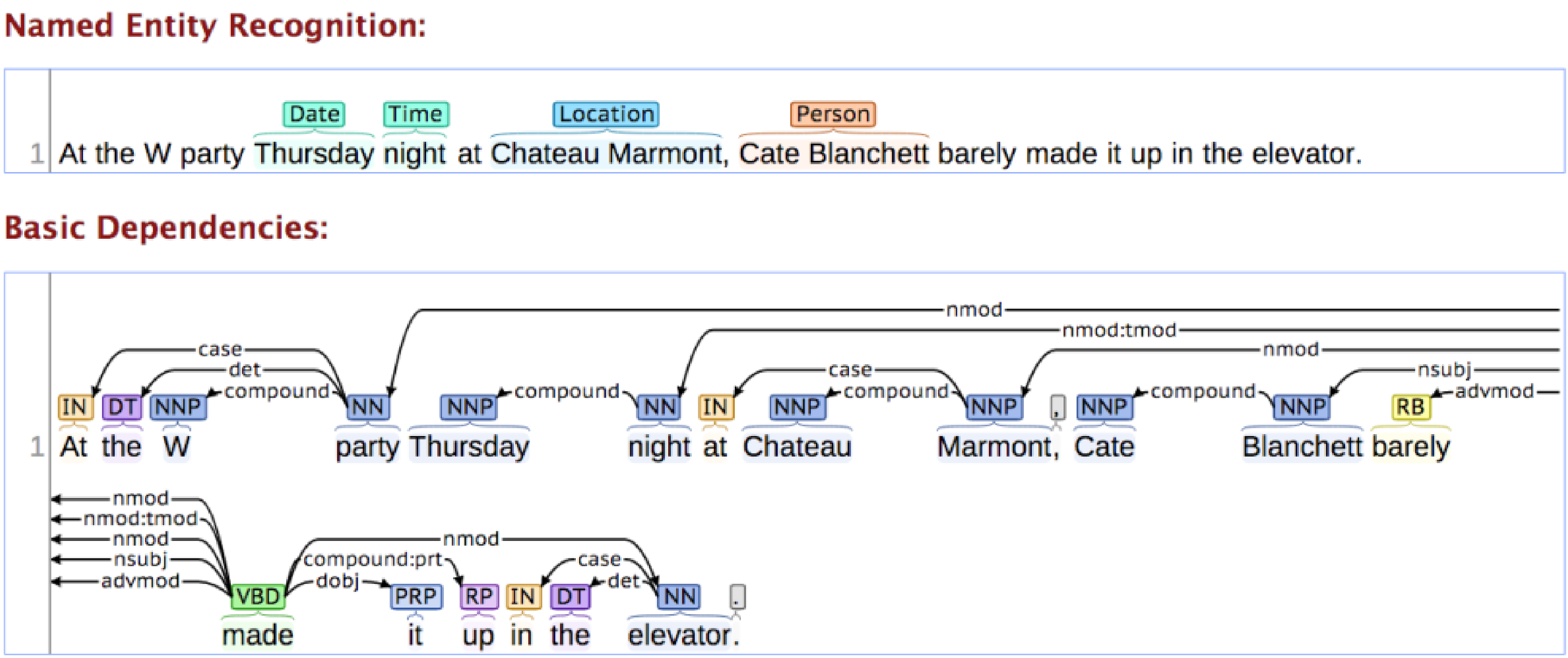
- Many named entities mentioned in tweets and other social media
- Extracting such entities (and also relations) enables us to pose interesting queries
- Interesting research question: what *is* an entity?
  - Weather warnings, disaster types, wind speeds…

Motivation for NER

- Many named entities mentioned in tweets and other social media
- Extracting such entities (and also relations) enables us to pose interesting queries
- Interesting research question: what *is* an entity?
- Good entity extraction proves crucial in *event extraction*, a much harder problem (covered later)

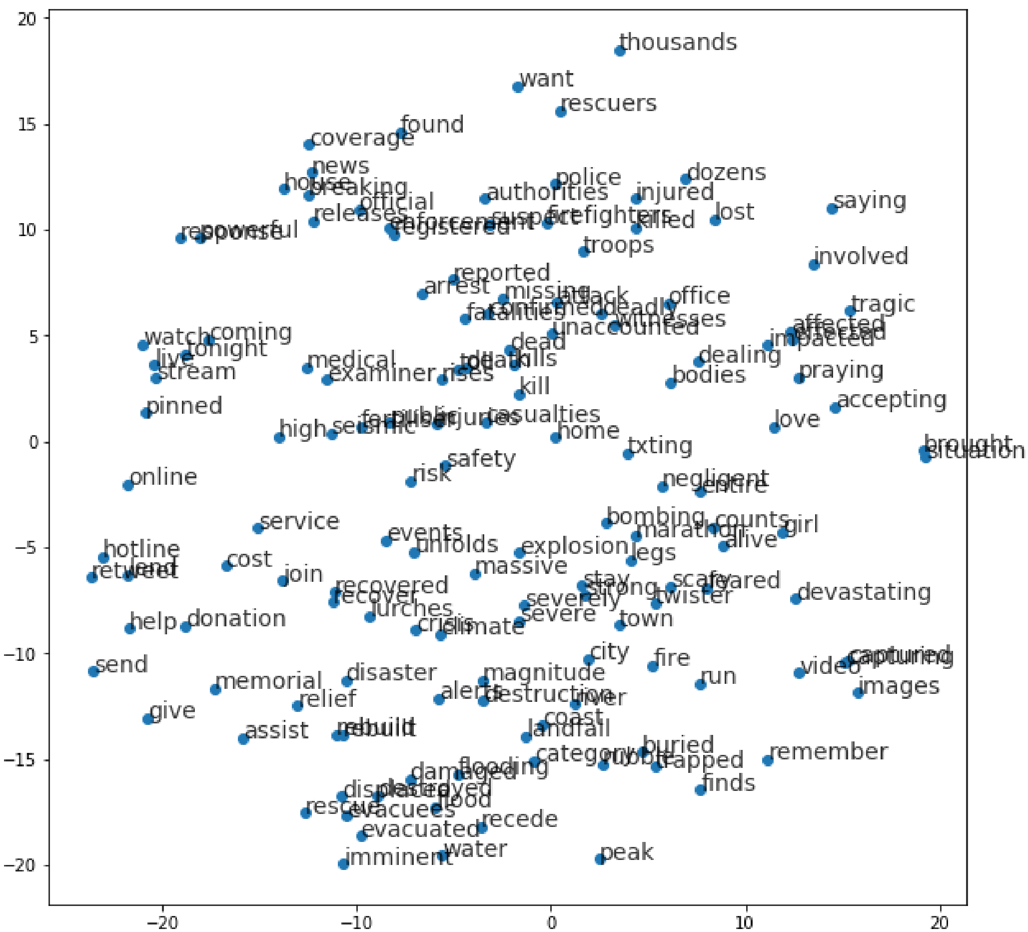
Classic NER Approach

- Till recently, most models framed the problem as 'sequence labeling' using techniques like Conditional Random Fields or (earlier) Hidden Markov Models



(Lafferty et al., 2001)

Embedding-based Models



- Feature engineering was an impediment to training robust and powerful CRFs
- Recently, word embeddings (and more complex sense-aware variants) have been used to address the problem

(Mikolov et al., 2013)

Embedding-based Models

- Feature engineering was an impediment to training robust and powerful CRFs
- Recently, word embeddings (and more complex sense-aware variants) have been used to address the problem

Demonstration of fastText over Twitter data

Powerful Tools Available

Deep Learning NLP with spaCy

spaCy

Stanford University

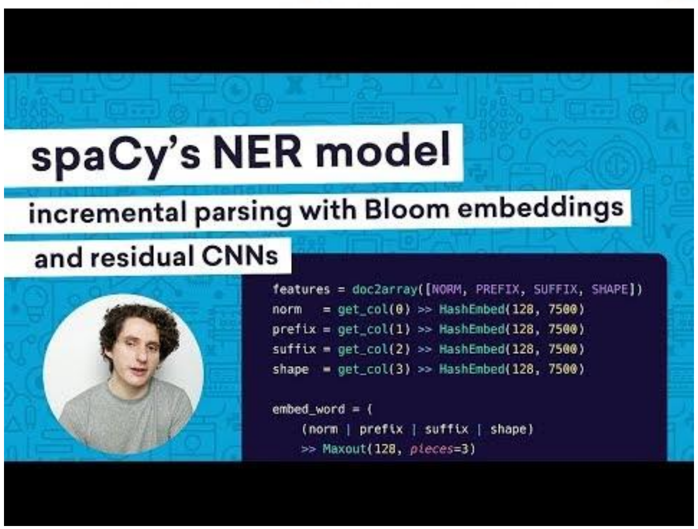
S

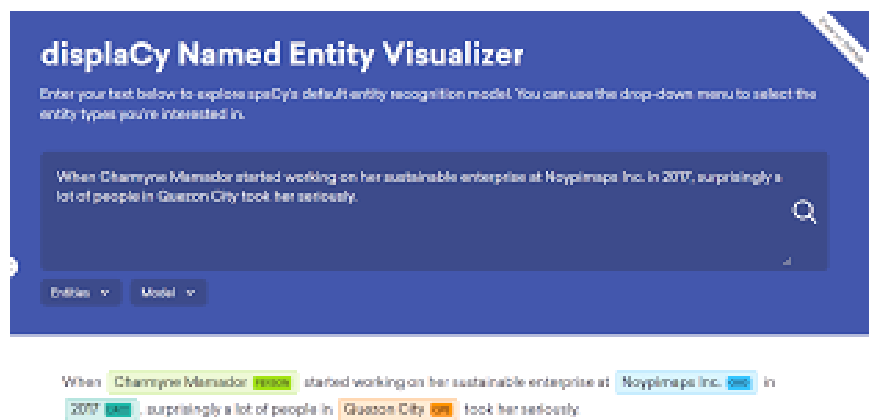
NLP

Natural Language Processing

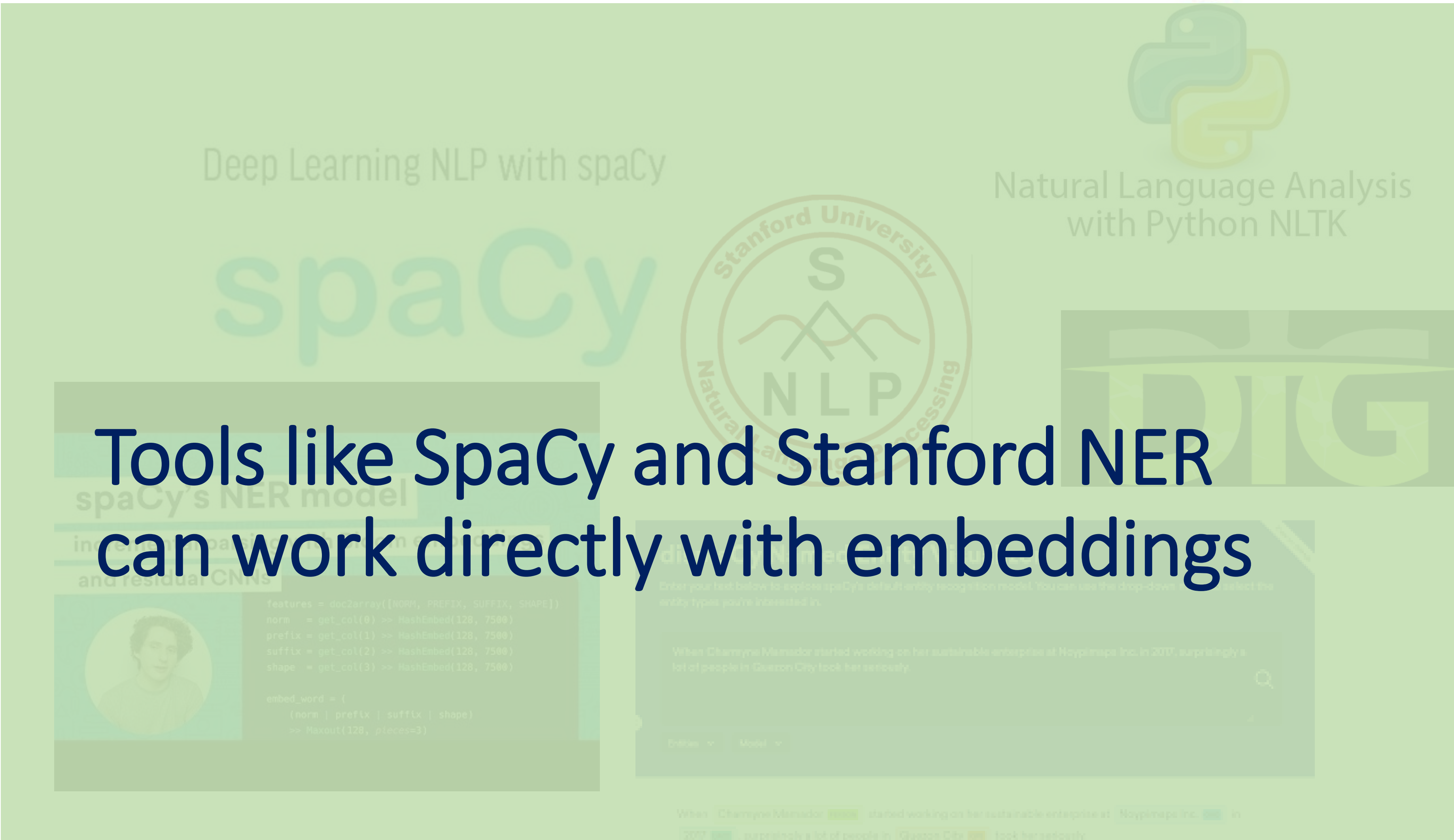
Natural Language Analysis
with Python NLTK

DIG





Powerful Tools Available



Tools like SpaCy and Stanford NER can work directly with embeddings

Powerful Tools Available

Demonstration of SpaCy

Are Off-the-shelf Tools Good Enough?

- **Example:** Stanford NER (off-the-shelf) vs. T-SEG (a Twitter-specific NER tool)
- P, R and F1 below stand for Precision, Recall and F1-Measure resp.

	P	R	F_1	F_1 inc.
Stanford NER	0.62	0.35	0.44	-
T-SEG(None)	0.71	0.57	0.63	43%
T-SEG(T-POS)	0.70	0.60	0.65	48%
T-SEG(T-POS, T-CHUNK)	0.71	0.61	0.66	50%
T-SEG(All Features)	0.73	0.61	0.67	52%

(Ritter et al., 2011)

Are Off-the-shelf Tools Good Enough?

- Example: Stanford NER (off-the-shelf) vs. T-SEG (a Twitter-specific NER tool)
- P, R and F1 below stand for Precision, Recall and F1-Measure res

Training Twitter-specific models and using Twitter-specific features offers significant performance advantages

	P	R	F_1	F_1 inc.
Stanford NER	0.62	0.35	0.44	-
T-SEG(None)	0.71	0.57	0.63	43%
T-SEG(T-POS)	0.70	0.60	0.65	48%
T-SEG(T-POS, T-CHUNK)	0.71	0.61	0.66	50%
T-SEG(All Features)	0.73	0.61	0.67	52%

(Ritter et al., 2011)

Twitter-specific NER Systems

Common Themes

- Best systems maximize 'signal' by leveraging joint contexts, distributional similarity, word embeddings and even URLs
- Geotagging tweets has emerged as its own 'mini-area' of research in the KDD, SW and WWW communities
- Performance is improving slowly, albeit still far from performance on traditional inputs like newswire

Impact

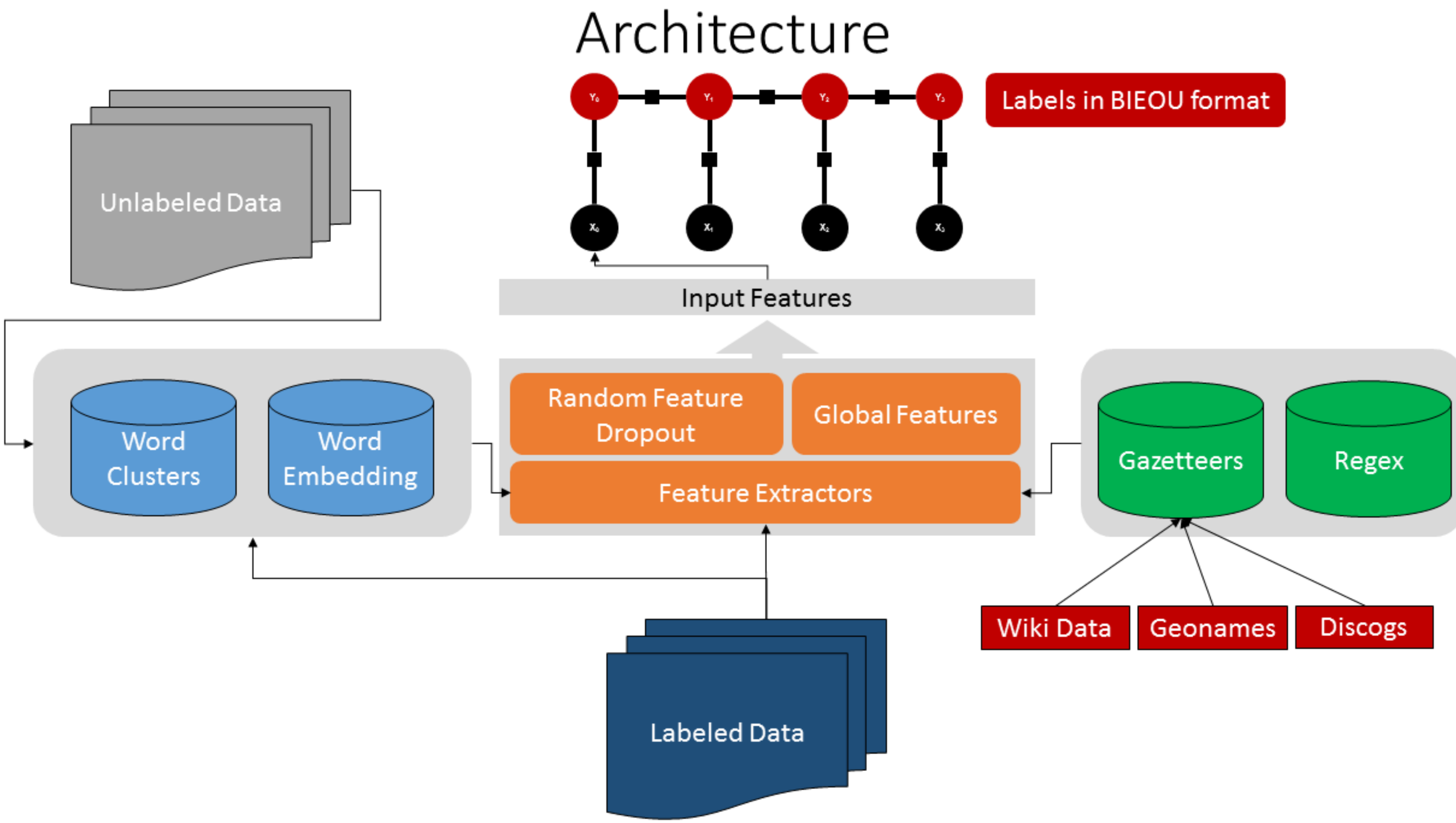
- Social-media specific NLP packages have emerged e.g., ArkNLP, T-SEG; workshops, shared competitions etc.
- Lots of research into how to parse irregular text, NLP methods have arguably become more robust as a result
- Spurred research on joint models, cross-domain entity linking

Classic System: T-SEG

- First system to (arguably) show that Twitter-specific NER far outperforms off-the-shelf state-of-the-art NERs
- Standard features e.g., POS tags, with some optimized for Twitter
  - Twitter-specific features include new tags for hashtags, retweets etc.
  - Showed results earlier
- In-domain training data i.e. *actual tweets*
  - Also used IRC chat data to supplement small training data
- Used distributional similarity to account for spelling variations,
  - Predated similar 'word embedding' techniques like fastText by many years (conceptually)!
  - Clusters words like 'tomarrow', 'tomm', 'tommarow', 'tommarrow'

(Ritter et al., 2011)

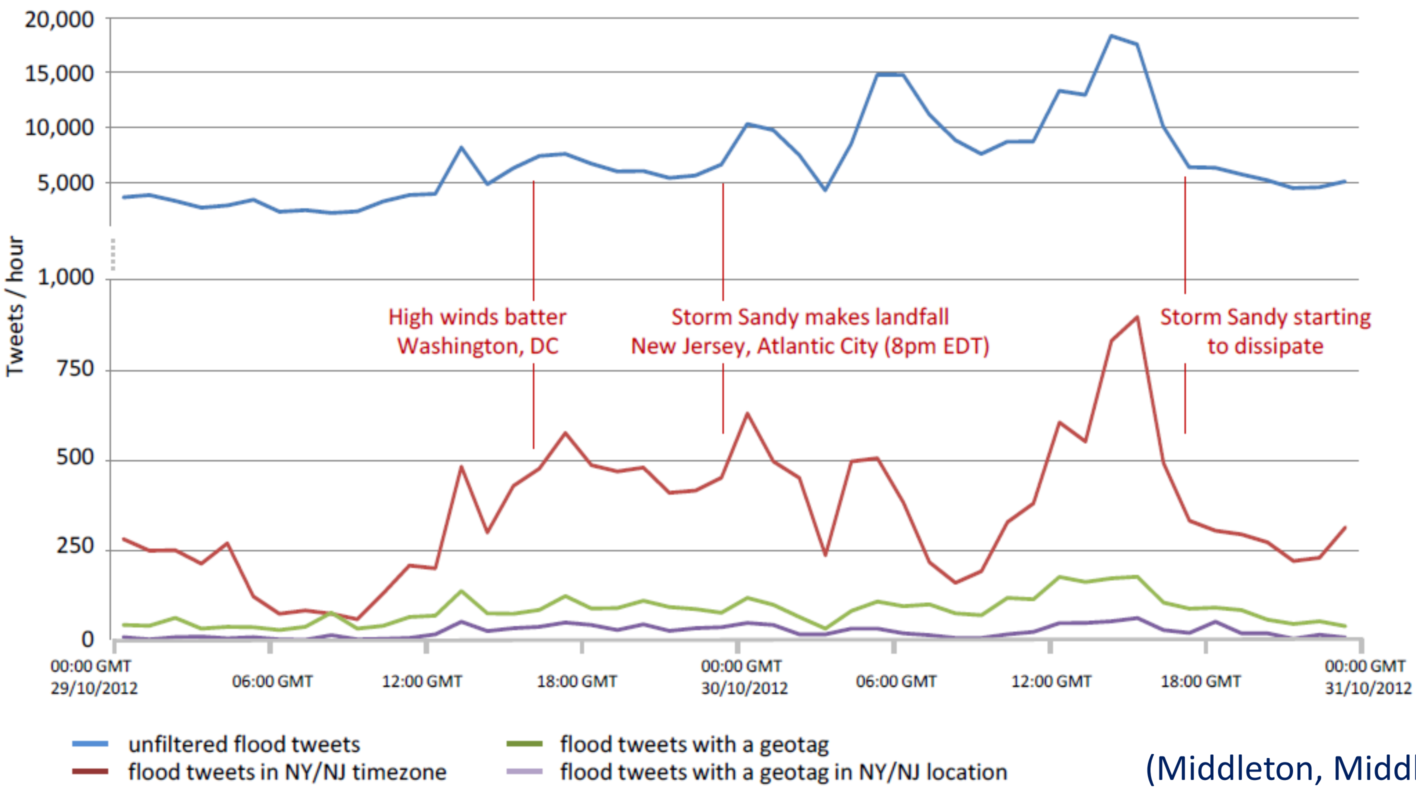
More Recent System: TwitterNER



(Mishra and Diesner, 2013)

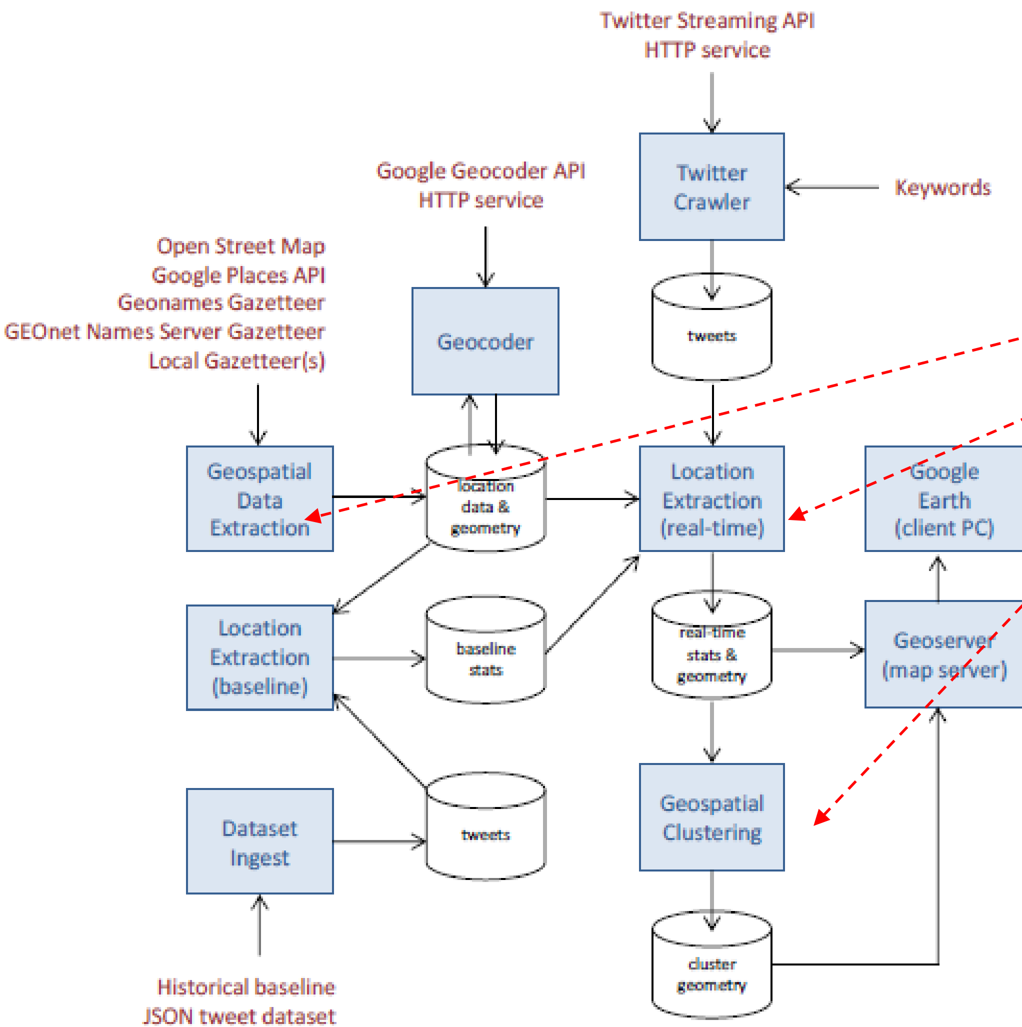
How to Further Improve?

- Models can be made more precise by treating each entity type (such as locations) individually i.e. train type-specific models
- In some instances, entities can be *inferred* despite not being explicitly present in the text (e.g., geotagging)



(Middleton, Middleton and Modafferi, 2014)

Example of crisis domain-specific geotagging system



Examples of domain-specific components

(Middleton, Middleton and Modafferi, 2014)

What happens after NER?

- What if the same entity got extracted **multiple times** in the text?
- What is the same entity got extracted **multiple times** in **multiple texts**?
- NER system can't tell that it is dealing with 'one' entity…treats every extraction as separate!

"I voted for Nader because he was most aligned with my values," she said.

Entity Linking/Resolution

- *Entity Resolution* is the problem of automatically determining when a pair of entities (extracted or otherwise) refers to the *same* underlying entity

DBpedia Spotlight

Alchemy (IBM)

Babelfy (BabelNet)

Text Razor NLP API

Aylien Text Analysis API

Open Research Issues

- Accuracy still low for social media (SM) NER
  - How to improve performance without increasing training annotations?
- How to work directly with noisy inputs (e.g., machine translated texts) and consume noisy NER outputs?
- NER for cross-domain and multi-lingual/non-English SM
  - Chinese social media (He and Sun, AAAI'17)

Open Research Issues

- How to leverage external contexts such as URLs in tweets, images, multi-modal signals, entity linking to sources like DBpedia...?
  - Can significantly enhance the ‘signal’ in the data e.g., see (Gattani et al., VLDB’13)
- How to combine NER and event identification/extraction models by leveraging joint context?
  - Promising work in this area e.g., (Vavliakis et al., DKE, 13)
- Novel applications and interfaces for crisis informatics pipelines

Summary

- Named Entity Recognition (NER) is an important problem in NLP and any situational awareness pipeline
- NER quality is much lower on Twitter data than 'ordinary' corpora like news or long text articles
  - Not known how tools currently perform on crisis-specific data
- State-of-the-art techniques make extensive use of embeddings and other creative uses of neural networks
- NER is only the first step, one must also perform co-reference resolution and entity linking!