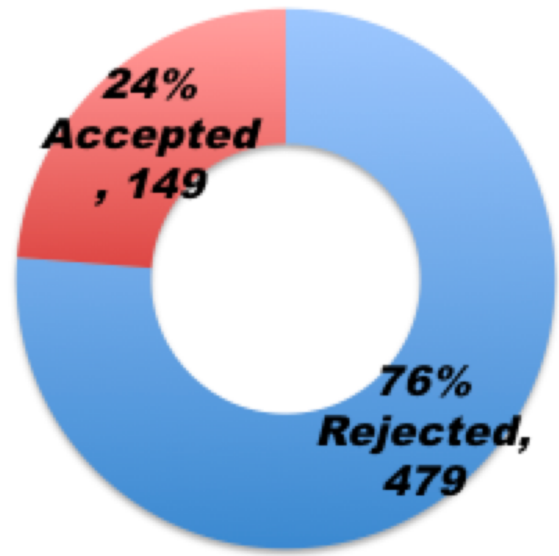
Named Entity Recognition in Tweets
- An Experimental Study

-- Presented By Haoyu Chen

Basic Info

Conference: Empirical Methods in Natural Language Processing (EMNLP)
Authors' institution : University of Washington, Seattle, WA
Time : 2011
Google Citation: 305

Accepted Rate -
24% at 2011 EMNLP



Outline

- Background
- Challenges
- Methodology & Evaluation
- Contributions

Review -- Named Entity Recognition (NER)

Jim bought 300 shares of Acme Corp. in 2006.

[Jim]$_{\text{Person}}$ bought 300 shares of [Acme Corp.]$_{\text{Organization}}$ in [2006]$_{\text{Time}}$.

Review -- BIO Encoding for NER

For example, we need name **person** and **date** entity, then define new tags.

Begin: *B-PERS, B-DATE*

-- Beginning of a mention of a person/date

Inside: *I-PERS, I-DATE*

-- Inside of a mention of a person/date

Outside: *O*

--outside of any mention of a named entity

<POS Tagging with restricted Tagset>

Challenges in Tweets

- Fresh Big Data

8,842 Tweets sent in 1 second

762,835,646

Tweets sent today

view how many in 1 second

Source: http://www.internetlivestats.com/twitter-statistics/

Challenges in Tweets

- Fresh Big Data

  => Out Of Vocabulary (OOV) , More Types of Entities
- Informal and Noisy Text

  => OOV, Uninformative Capitalization

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...

Table 1: Examples of noisy text in tweets.

Challenges in Tweets

- Fresh Big Data

  => Out Of Vocabulary (OOV) , More Types of Entities

- Informal and Noisy Text

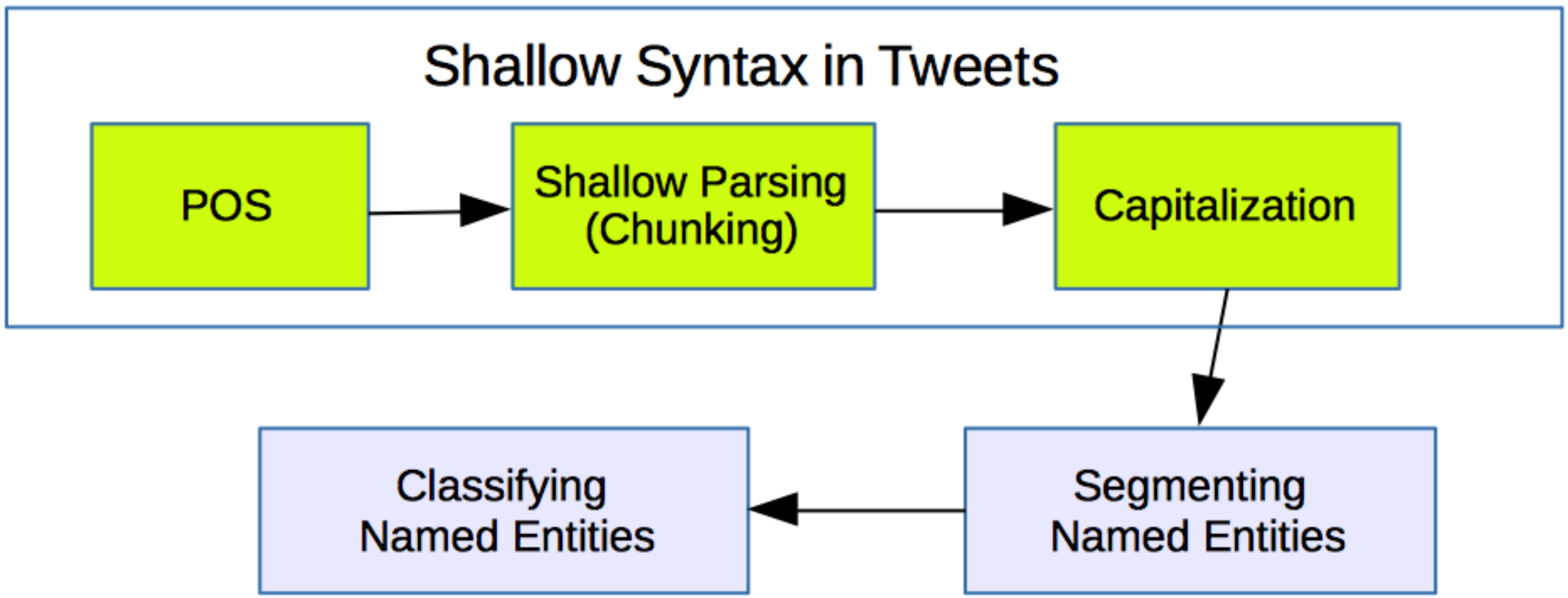
  => OOV, Uninformative Capitalization

- 140 characters Limit

  => Lack of Context

KKTNY in 45min..........

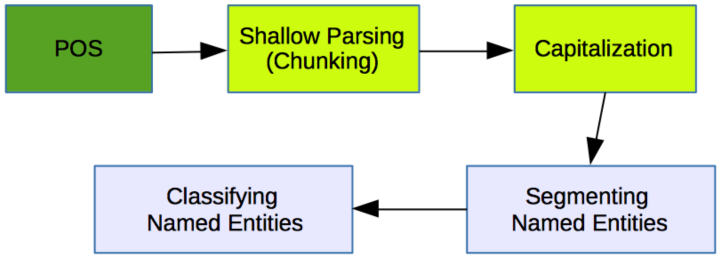
Methodology -- System Flow Chart



Traditional POS Tagging

Data Set,	Baseline
Brown Corpus,	0.90
Tweets,	0.76

Data Set,	State-Of-Art
WSJ,	0.97
Tweets,	0.80



Improvements on POS

- Apply hierarchical clustering (Brown Clustering) on 52 million tweets to capture lexical variations.

‘2m’, ‘2ma’, ‘2mar’, ‘2mara’, ‘2maro’, ‘2marrow’, ‘2mor’, ‘2mora’, ‘2moro’, ‘2morow’, ‘2morr’, ‘2morro’, ‘2morrow’, ‘2moz’, ‘2mr’, ‘2mro’, ‘2mrrw’, ‘2mrw’, ‘2mw’, ‘tmmrw’, ‘tmo’, ‘tmoro’, ‘tmorrow’, ‘tmoz’, ‘tmr’, ‘tmro’, ‘tmrow’, ‘tmrrow’, ‘tmrrw’, ‘tmrw’, ‘tmrww’, ‘tmw’, ‘tomaro’, ‘tomarow’, ‘tomarro’, ‘tomarrow’, ‘tomm’, ‘tommarow’, ‘tommarrow’, ‘tommoro’, ‘tommorow’, ‘tommorrow’, ‘tommorw’, ‘tommrow’, ‘tomo’, ‘tomolo’, ‘tomoro’, ‘tomorow’, ‘tomorro’, ‘tomorrw’, ‘tomoz’, ‘tomrw’, ‘tomz’

Improvements on POS

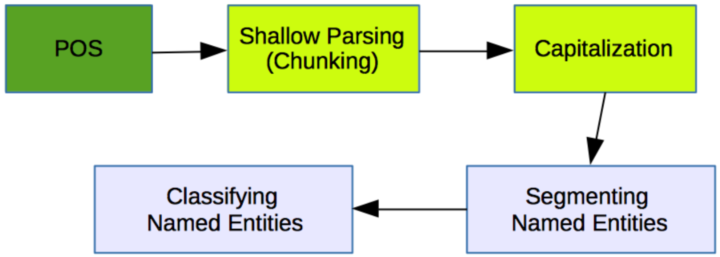
- Apply hierarchical clustering (Brown Clustering) on 52 million tweets to capture lexical variations.
- Conditional Random Fields is used to get the help of the context.

Simple Example: I will be there toma.

I will be there [tomorrow **Adv**]. ✓

I will be there [tomato **noun**]. X

Improvements on POS



- Apply hierarchical clustering (Brown Clustering) to capture lexical variations.
- Conditional Random Fields is used to get the help of context .

- Add new tags, such as urls, #hashtags,@usernames, and retweets. (100% accuracy)
- To overcome difference in style and vocabulary, they manually annotated 800 tweets as in-domain training data.
- Incorporate out-domain training data, such as IRC.

Evaluation on POS (4-fold validation)

	Accuracy	Error Reduction
Majority Baseline (NN)	0.189	-
Word's Most Frequent Tag	0.760	-
Stanford POS Tagger	0.801	-
T-POS(PTB)	0.813	6%
T-POS(Twitter)	0.853	26%
T-POS(IRC + PTB)	0.869	34%
T-POS(IRC + Twitter)	0.870	35%
T-POS(PTB + Twitter)	0.873	36%
T-POS(PTB + IRC + Twitter)	0.883	41%

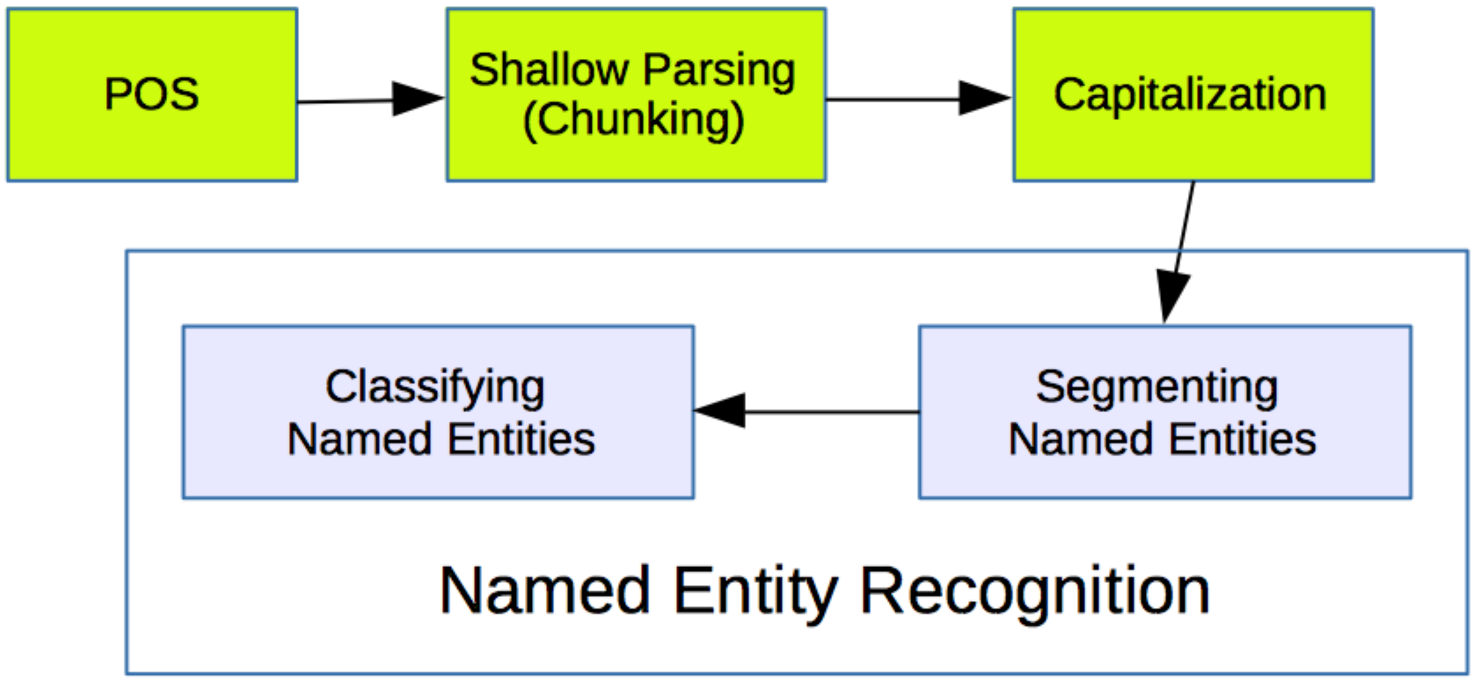
Table 2: POS tagging performance on tweets. By training on in-domain labeled data, in addition to annotated IRC chat data, we obtain a 41% reduction in error over the Stanford POS tagger.

Evaluation--Shallow Parsing (Chunking)

	Accuracy	Error Reduction
Majority Baseline (B-NP)	0.266	-
OpenNLP	0.839	-
T-CHUNK(CoNLL)	0.854	9%
T-CHUNK(Twitter)	0.867	17%
T-CHUNK(CoNLL + Twitter)	0.875	22%

Table 4: Token-Level accuracy at shallow parsing tweets. We compare against the OpenNLP chunker as a baseline.

Methodology -- System Flow Chart



Segmenting Named Entities

- Larger annotated dataset to effectively learn a model of named entities - Randomly sampled 2400 tweets.
- BOI encoding for representing segmentation
- Dictionaries included a set of type lists gathered from Freebase
- Use the result of shallow syntax as input

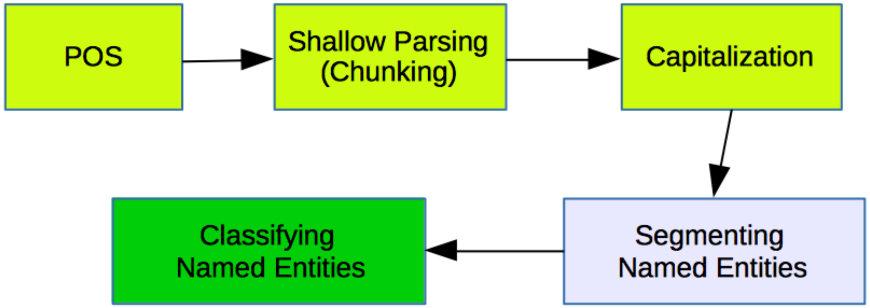
Note: Freebase is an online collection of structured data and Google's Knowledge Graph is powered in part by it

Evaluation--Segmenting Named Entities

	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%

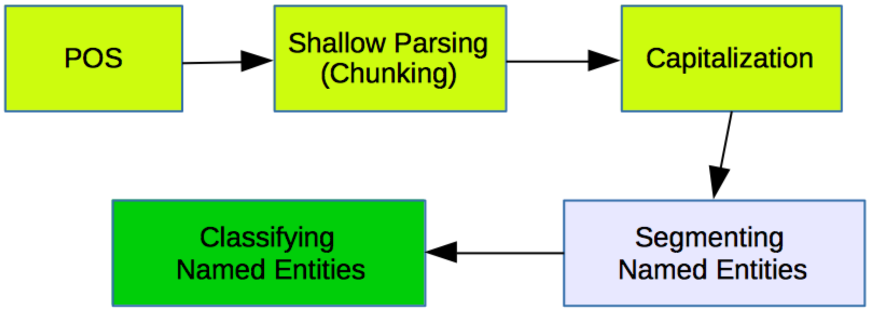
Classifying Named Entities

- Distinctive entities => Big training dataset ?
- Lack of context => Out-domain knowledge
-
- **Solution:** Leverage large lists of entities gathered from Freebase as a source of distant supervision
- **Benefit:** Allow use of large amount of unlabeled data in learning



Classifying Named Entities

- 30% of entities on Twitter are out of Freebase dictionary while 35% of entities on Twitter has multiple meaning types in Freebase.
- **Enhanced Solution:**Topic Model is to discover the hidden thematic structure in docs, and Latent Dirichlet allocation(LDA) is one of the mosted used topic model.
- **Basic Idea:**Every doc is made of multiple topics. The words in the documents are generated from those multiple topics.



Evaluation--Classifying Named Entities

System	P	R	F_1
Majority Baseline	0.30	0.30	0.30
Freebase Baseline	0.85	0.24	0.38
Supervised Baseline	0.45	0.44	0.45
DL-Cotrain	0.54	0.51	0.53
LabeledLDA	0.72	0.60	0.66

Table 8: Named Entity Classification performance on the 10 types. Assumes segmentation is given as in (Collins and Singer, 1999), and (Elsner et al., 2009).

Contributions

- Design and implement a **complete system for Named Entity Recognition (NER) in Tweets**. By optimizing each steps of NER system, it shows a substantially improvement on performance.

- It introduces a new approach to classify named entity by applying distant supervision with Topic Models (), which is able to **train large amount of unlabeled dataset.**

Thank you!

Any Questions?

Application Demo