Named Entity Recognition in Tweets : An Experimental Study

Alan Ritter, Sam Clark, Mausam and Oren Etzioni

Outline

- About Tweets
- Named entity recognition
- NER Pipeline
  - T-POS
  - T-CHUNK
  - T-CAP
  - T-SEG
  - T-NER
- Experiments and Results
- Related Work
- Conclusions

What is a tweet ??

Short status messages from users.

Maximum of 140 characters per message.

Government confirms blast n nuclear plants n japan...don’t knw wht s gona happen nw...

Named Entity Recognition ??

Companies, products, brands, people, locations etc..

Yess! Yess! Its official Nintendo announced today that theyWill release the Nintendo 3DS in north America march 27 for $250

Why NER So Difficult ?

- Plethora of distinctive named entity types but infrequent.
- 140 character limit hence lack of context.

Part of Speech Tagging

- Assigning tokens in the text their corresponding part of speech tags like NN, VB, NNS.

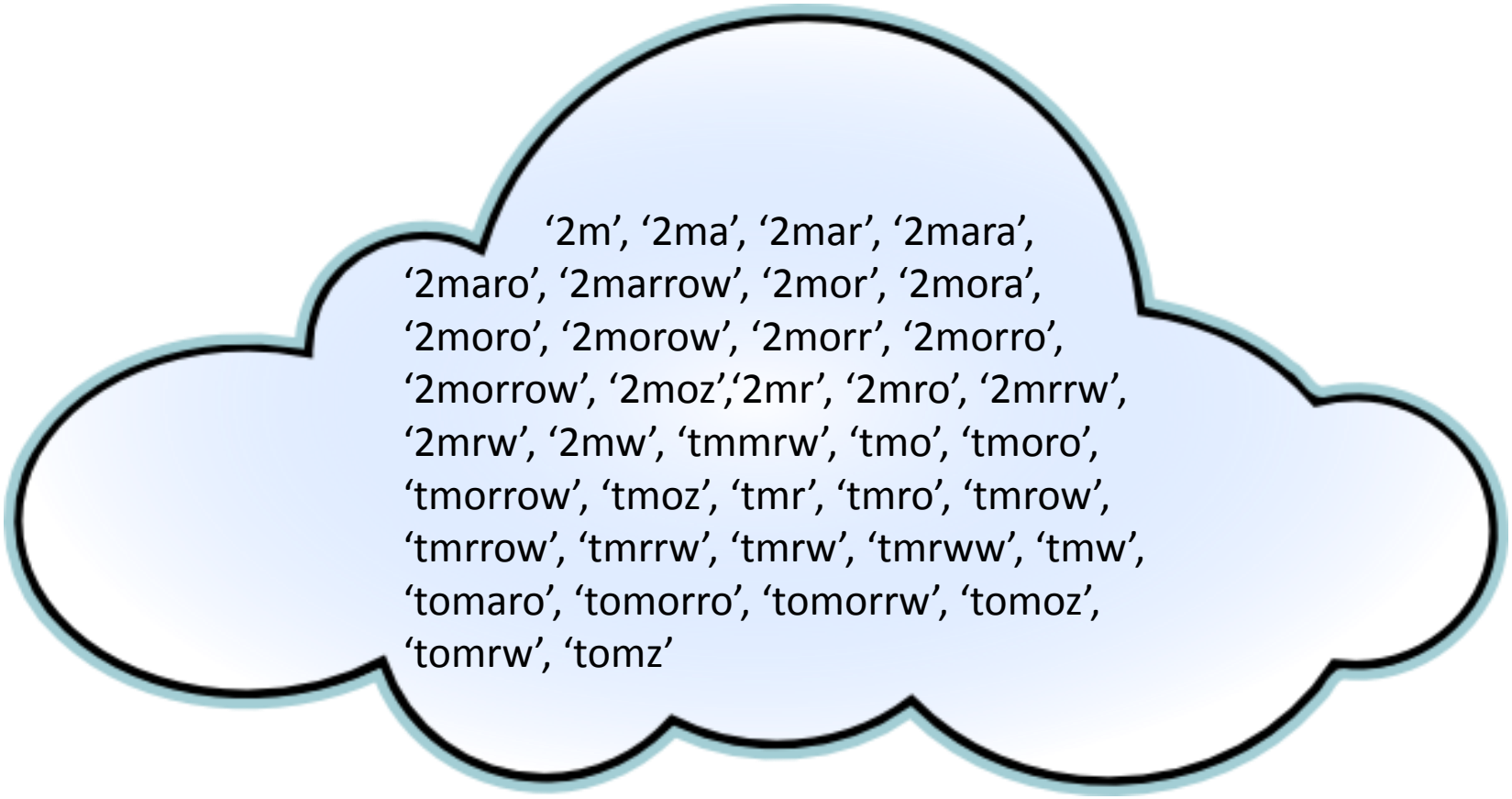
[***NP*** He] [***VPZ*** reckons] [***DT*** the] [***JJ*** current] [***NN*** account]
[***NN*** deficit] [***MD*** will] [***VB*** narrow] [***TO*** to] [***RB*** only]
[**#** #] [***CD*** 1.8] [***CD*** billion] [***IN*** in] [***NNP*** September] [**.** .]

T-POS

- Manually annotated 800 tweets (~ 16K tokens).
- Tags Used: Penn Treebank
- New Tags: retweets, @usernames, #hashtags, and urls.
- Clustering to deal with OOV words
  - Heirarchical clustering using Jcluster.
  - 52 million tweets used

- *Conditional Random Fields (CRF)* for sequential learning.

Clustering example



Chunking

- Identifying phrases such as noun phrases, verb phrases and prepositional phrases in the text.

[***NP*** He] [***VP*** reckons] [***NP*** the current account deficit] [***VP*** will narrow] [***PP*** to] [***NP*** only # 1.8 billion] [***PP*** in] [***NP*** September] .

T-Chunk

- Annotate same set of 800 tweets.
- Used tags from CoNLL shared task.
- POS features from T-POS.
- 16K tokens of in-domain training data. 210K tokens newswire text from CoNLL dataset.
- Used CRF again for inference and learning.

T-Chunk Results

	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%

T-CAP

- Predicting whether capitalization informative or not.
- Manually labeled 800 tweets as informative or not.

- Criteria Used:
  - Non Entity words beginning with capitalization and also not the start of the sentence.
  - Entities not beginning with capitalization.

- *Support Vector Machines* for Classification

Named Entity Recognition in tweets T-NER

- A Result from Stanford Named Entity Recognizer

[Yess]$_{ORG}$! [Yess]$_{ORG}$! Its official
[Nintendo$_{]LOC}$ announced today that they
Will release the [Nintendo]$_{ORG}$ 3DS in north
[America]$_{LOC}$ march 27 for $250

Consider classification and segmentation as two different tasks

Segmenting Named Entities T-SEG

- Capitalization less informative
  - Need for in-domain data relying less on capitalizations
  - Also should be able to use features from T-CAP

- Models NER as a sequence labeling task using IOB encoding.
  - Each word either begins, is inside or is outside of a named entity.
  - For example:

    The morning flight from Denver has arrived.

    B_NP I_NP I_NP O_NP B_NP O_NP O_NP

- Use outputs from T_POS, T-CHUNK and T_CAP

Classifying Named Entities

- Challenges in Classification
  - Twitter containing many distinctive and infrequent entity types
  - Collecting training data a difficult task

  - **For Example**, Consider

    KKTNY in 45 min ....

  - Without any priori difficult to identify which entity type KKTNY refers to.
  - Make use of contextual information like co-occurs with *watching* and *premieres* in other contexts

Distance Supervision using Topic modeling

- Entity in data associated with bag of words in the contextual window
  - Each bag of words associated with a Multinomial distribution of topics θ_e
  - Each topic associated with a distribution over words multinomial β_t
- One to One mapping between topics and Freebase type dictionaries
  - Constrains θ_e over set of possible types FB[e]
  - For example,

    θ_{Amazon} constrained over COMPANY and LOCATION

    θ_{Apple} constrained over COMPANY and FOOD
- If entities are absent in Freebase leave their θ_e unconstrained.

EXPERIMENTS

Classification Experiments

- Annotated 2400 tweets
- Tags Used:

PERSON
GEO-LOCATION,
COMPANY,
PRODUCT,
FACILITY,
TV-SHOW,
MOVIE,
SPORTSTEAM,
BAND
OTHER

Classification Experiments
Training

- Run T-SEG on 60M tweets to extract entities (100 or more times)
- Results in 23,651 distinct entity strings
- Collect words from contextual window of size 3
- Gibbs sampling for 1000 iterations to estimate θ_e and β_t

Type List produced by LabeledLDA

Type	Top 20 Entities not found in Freebase dictionaries
PRODUCT	nintendo ds lite, apple ipod, generation black, ipod nano, apple iphone, gb black, xperia, ipods, verizon media, mac app store, kde, hd video, nokia n8, ipads, iphone/ipod, galaxy tab, samsung galaxy, playstation portable, nintendo ds, vpn
TV-SHOW	pretty little, american skins, nof, order svu, greys, kktny, rhobh, parks & recreation, parks & rec, dawson 's creek, big fat gypsy weddings, big fat gypsy wedding, winter wipeout, jersey shores, idiot abroad, royle, jerseyshore, mr . sunshine, hawaii five-0, new jersey shore
FACILITY	voodoo lounge, grand ballroom, crash mansion, sullivan hall, memorial union, rogers arena, rockwood music hall, amway center, el mocambo, madison square, bridgestone arena, cat club, le poisson rouge, bryant park, mandalay bay, broadway bar, ritz carlton, mgm grand, olympia theatre, consol energy center

T-CLASS Classification Results

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

- *Majority Baseline:* most frequent label (PERSON)
- *Freebase Baseline:* makes prediction only if entity unambiguous.
- *Supervised Baseline:* MaxEnt classifier.
- *Co-Training Algorithm:* also uses Freebase
- *T-CLASS outperforms baselines and achieves 25% increase in F1 score over Cotrain*

F1 classification scores comparisons

PERSON, LOCATION, ORGANIZATION

Type	LL	FB	CT	SP	N
PERSON	0.82	0.48	0.65	0.83	436
LOCATION	0.74	0.21	0.55	0.67	372
ORGANIZATION	0.66	0.52	0.55	0.31	319
overall	0.75	0.39	0.59	0.49	1127

F1 Scores for all types

Type	LL	FB	CT	SP	N
PERSON	0.82	0.48	0.65	0.86	436
GEO-LOC	0.77	0.23	0.60	0.51	269
COMPANY	0.71	0.66	0.50	0.29	162
FACILITY	0.37	0.07	0.14	0.34	103
PRODUCT	0.53	0.34	0.40	0.07	91
BAND	0.44	0.40	0.42	0.01	54
SPORTSTEAM	0.53	0.11	0.27	0.06	51
MOVIE	0.54	0.65	0.54	0.05	34
TV-SHOW	0.59	0.31	0.43	0.01	31
OTHER	0.52	0.14	0.40	0.23	219
overall	0.66	0.38	0.53	0.45	1450

Results some more…

	P	R	F_1
DL-Cotrain-entity	0.47	0.45	0.46
DL-Cotrain-mention	0.54	0.51	**0.53**
LabeledLDA-entity	0.73	0.60	**0.66**
LabeledLDA-mention	0.57	0.52	0.54

LabeledLDA vs DL-Cotrain
Grouping unlabeled data by entities vs mentions

System	P	R	F_1
COTRAIN-NER (10 types)	0.55	0.33	0.41
T-NER(10 types)	0.65	0.42	**0.51**
COTRAIN-NER (PLO)	0.57	0.42	0.49
T-NER(PLO)	0.73	0.49	**0.59**
Stanford NER (PLO)	0.30	0.27	0.29

Performance at predicting both segmentation and classification

Related Work

- Locke and Martin (2009) train classifier handling PERSON, LOCATION and ORGANIZATION.
- Liu at. al. (2011) PRODUCTS, K-Nearest Neighbors.
- Gimpell at. el. (2011) POS tagger using coarse-grained tags.
- Benson at. el. (2011) artists and venues from musical performances.
- Finin at el. (2010) Amazon's Mechanical Turks for NER in tweets.
- Minkov et. El.(2005) person name recognizer in emails.
- Singh at. El. (2010) NER in text advertisements.

Thank You !

- Questions ??