Social Media & Text Analysis

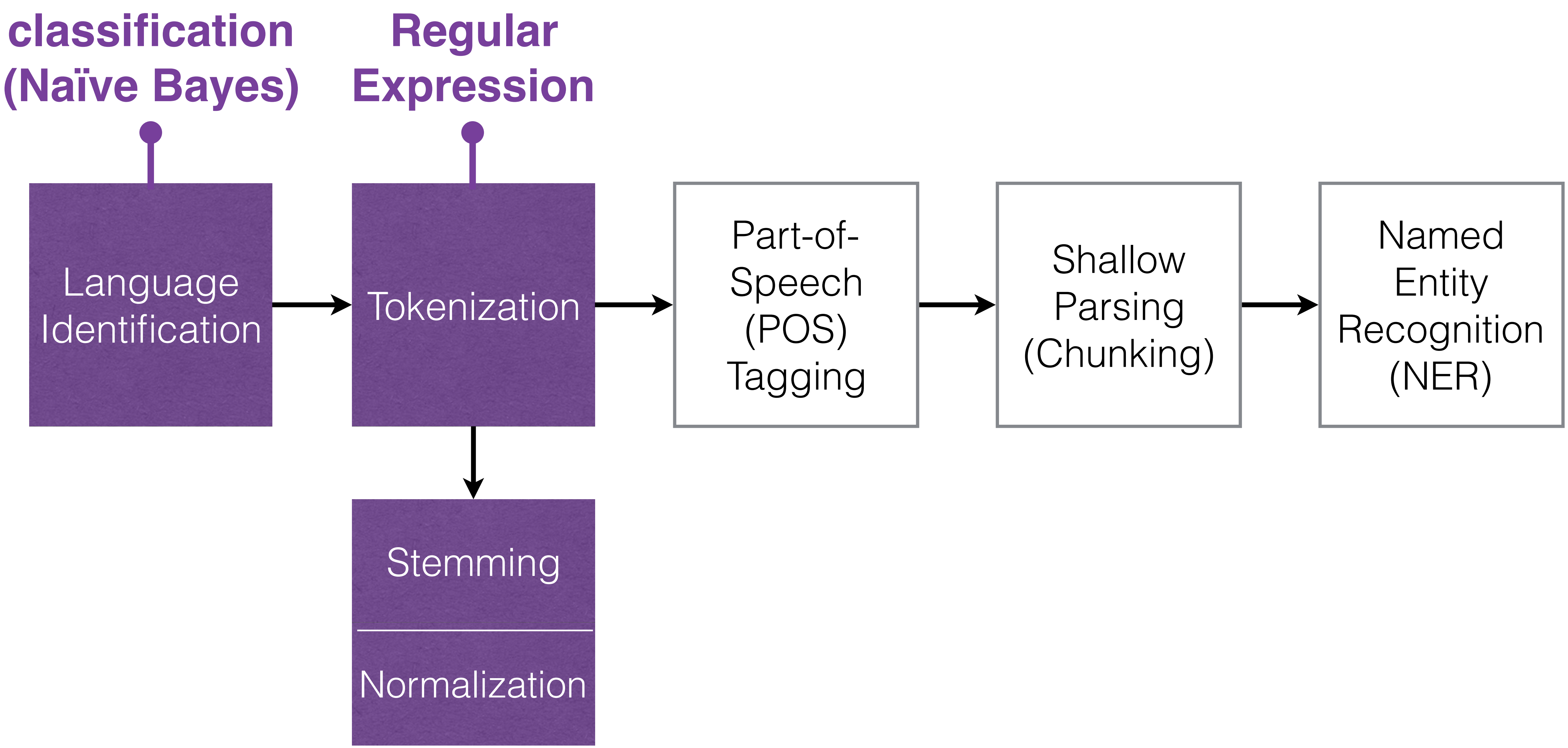
lecture 5 - POS/NE Tagging

CSE 5539-0010 Ohio State University

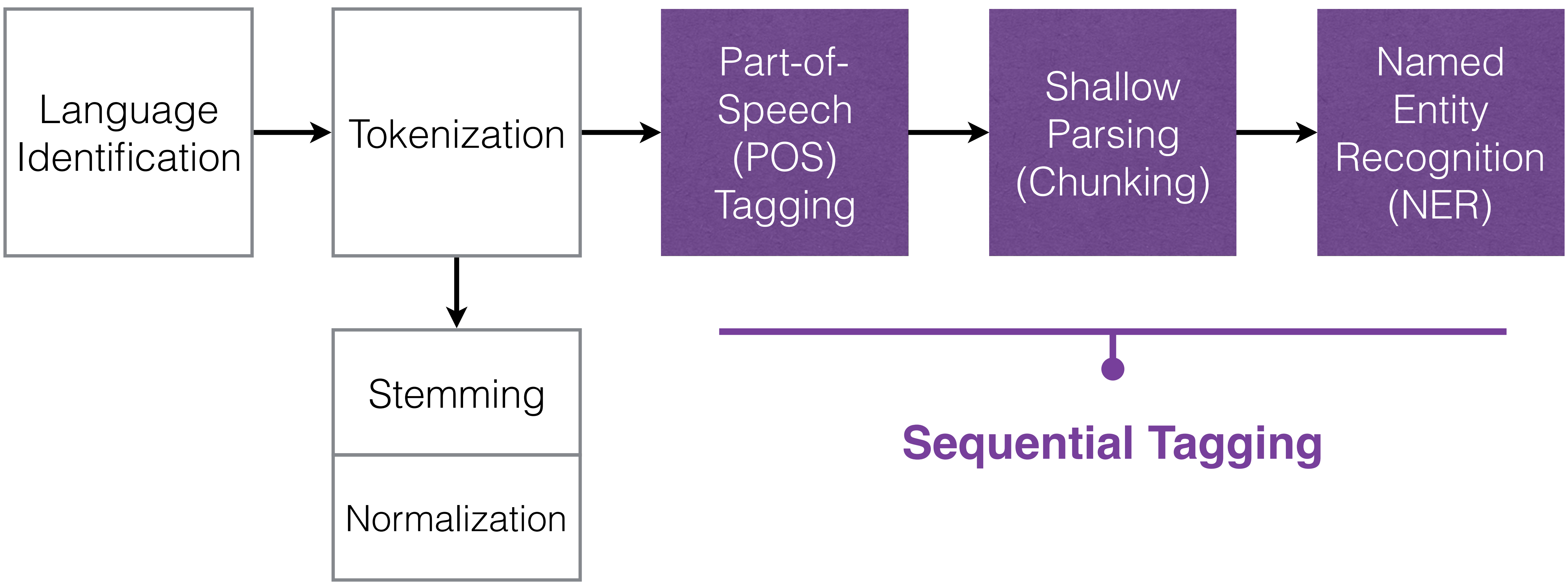
Instructor: Alan Ritter

Website: socialmedia-class.org

NLP Pipeline (summary so far)



NLP Pipeline (next)



Challenge: Natural Language Processing Breaks

Sohaib Athar
@ReallyVirtual

Helicopter hovering above Abbottabad at 1AM (is a rare event).

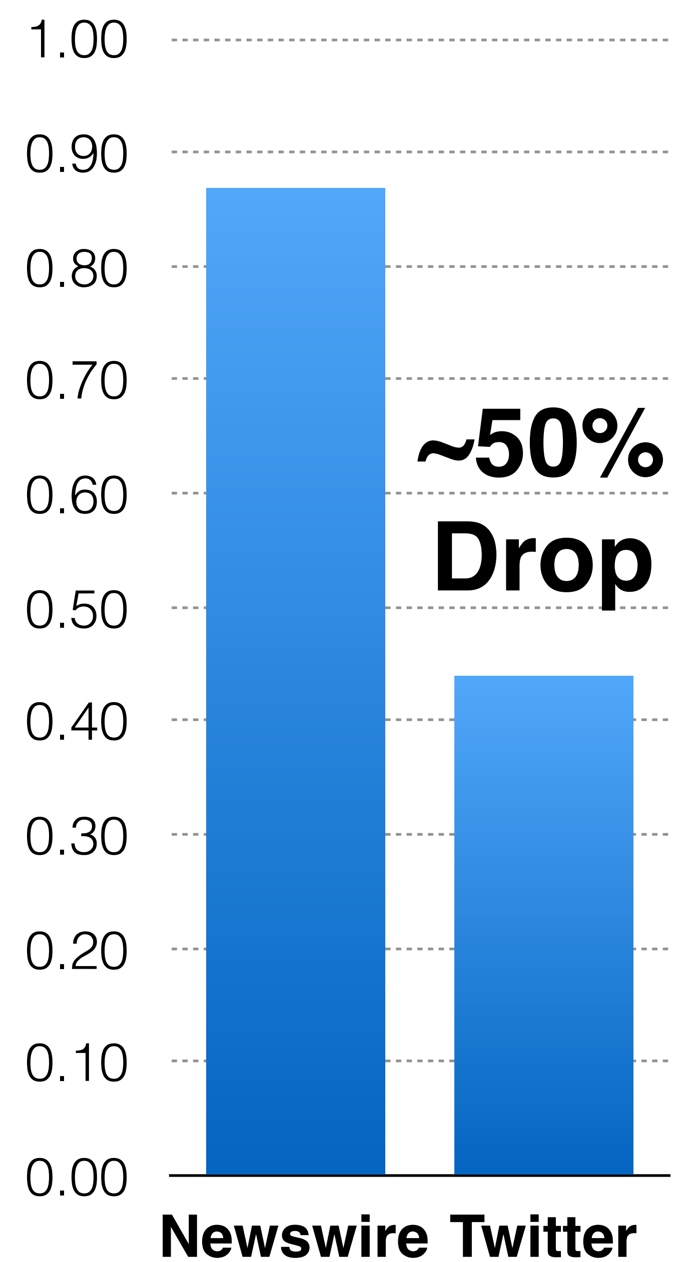
Keith Urbahn
@keithurbahn

So I'm told by a reputable person they have killed Osama Bin Laden. Hot damn.

Challenge: Natural Language Processing Breaks

Challenge: Natural Language Processing Breaks

Stanford NER:



Part-of-Speech (POS) Tagging

Cant	MD
wait	VB
for	IN
the	DT
ravens	NNP
game	NN
tomorrow	NN
…	:
go	VB
ray	NNP
rice	NNP
!!!!!!!	.

Penn Treebank POS Tags

1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition/subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol (mathematical or scientific)
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund/present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd ps. sing. present
32.	VBZ	Verb, 3rd ps. sing. present
33.	WDT	*wh*-determiner
34.	WP	*wh*-pronoun
35.	WP$	Possessive *wh*-pronoun
36.	WRB	*wh*-adverb
37.	#	Pound sign
38.	$	Dollar sign
39.	.	Sentence-final punctuation
40.	,	Comma
41.	:	Colon, semi-colon
42.	(	Left bracket character
43.	)	Right bracket character
44.	"	Straight double quote
45.	‘	Left open single quote
46.	“	Left open double quote
47.	’	Right close single quote
48.	”	Right close double quote

Part-of-Speech (POS) Tagging

- Words often have more than one POS:
  - The back door = JJ
  - On my back = NN
  - Win the voters back = RB
  - Promised to back the bill = VB

- POS tagging problem is to determine the POS tag for a particular instance of a word.

Source: adapted from Chris Manning

Twitter-specific Tags

- #hashtag
- @metion
- url
- email address
- emoticon
- discourse marker
- symbols
- …

Source: Gimpel et al. "Part-of-Speech Tagging for Twitter : Annotation, Features, and Experiments" ACL 2011

Noisy Text: Challenges

- Lexical Variation (misspellings, abbreviations)
  - `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‘
- Unreliable Capitalization
  - “The Hobbit has FINALLY started filming! I cannot wait!”
- Unique Grammar
  - “watchng american dad.”

Chunking

Cant	VP
wait	
for	PP
the	NP
ravens	
game	
tomorrow	NP
…	
go	VP
ray	NP
rice	
!!!!!!!	

Chunking

- recovering phrases constructed by the part-of-speech tags
- a.k.a shallow (partial) parsing:
  - full parsing is expensive, and is not very robust
  - partial parsing can be much faster, more robust, yet sufficient for many applications
  - useful as input (features) for named entity recognition or full parser

Named Entity Recognition(NER)

Cant	
wait	
for	
the	
ravens	ORG
game	
tomorrow	
…	
go	
ray	PER
rice	
!!!!!!!	.

ORG: organization

PER: person

LOC: location

NER: Basic Classes

Cant	
wait	
for	
the	
ravens	ORG
game	
tomorrow	
…	
go	
ray	PER
rice	
!!!!!!!	.

ORG: organization

PER: person

LOC: location

Noisy Text: NLP breaks

POS:

NNP/ Yess ./ ! NNP/ Yess ./ ! PRP$/ Its JJ/ official NNP/ Nintendo VBD/ announced NN/ today IN/ that PRP/ they MD/ Will VB/ release DT/ the NNP/ Nintendo NN/ 3DS IN/ in RB/ north NNP/ America NN/ march CD/ 27 IN/ for NN/ $250

Chunk:

[NP Yess] ! [NP Yess] ! [NP Its offcial Nintendo] [VP announced]
[NP today] [SBAR that] [NP they] [VP Will release] [NP the Nintendo 3DS]
[PP in] [NP north America march] [NP 27] [PP for] [NP $250]

NER:

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

Noisy Text: NLP breaks

POS:

NNP/ Yess ./ ! NNP/ Yess ./ ! PRP$/ Its JJ/ official NNP/ Nintendo VBD/ announced NN/ today IN/ that PRP/ they MD/ Will VB/ release DT/ the NNP/ Nintendo NN/ 3DS IN/ in RB/ north NNP/ America NN/ march CD/ 27 IN/ for NN/ $250

Chunk:

[NP Yess] ! [NP Yess] ! [NP Its offcial Nintendo] [VP announced]
[NP today] [SBAR that] [NP they] [VP Will release] [NP the Nintendo 3DS]
[PP in] [NP north America march] [NP 27] [PP for] [NP $250]

NER:

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

Noisy Text: NLP breaks

POS:

NNP/ Yess ./ ! NNP/ Yess ./ ! PRP$/ Its JJ/ official NNP/ Nintendo VBD/ announced NN/ today IN/ that PRP/ they MD/ Will VB/ release DT/ the NNP/ Nintendo NN/ 3DS IN/ in RB/ north NNP/ America NN/ march CD/ 27 IN/ for NN/ $250

Chunk:

[NP Yess] ! [NP Yess] ! [NP Its offcial Nintendo] [VP announced] [NP today] [SBAR that] [NP they] [VP Will release] [NP the Nintendo 3DS] [PP in] [NP north America march] [NP 27] [PP for] [NP $250]

NER:

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

Noisy Text: NLP breaks

POS:

NNP/ Yess ./ ! NNP/ Yess ./ ! PRP$/ Its JJ/ official NNP/ Nintendo VBD/ announced NN/ today IN/ that PRP/ they MD/ Will VB/ release DT/ the NNP/ Nintendo NN/ 3DS IN/ in RB/ north NNP/ America NN/ march CD/ 27 IN/ for NN/ $250

Chunk:

[NP Yess] ! [NP Yess] ! [NP Its offcial Nintendo] [VP announced] [NP today] [SBAR that] [NP they] [VP Will release] [NP the Nintendo 3DS] [PP in] [NP north America march] [NP 27] [PP for] [NP $250]

NER:

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

Noisy Text: NLP breaks

POS:

NNP/ Yess ./ ! NNP/ Yess ./ ! PRP$/ Its JJ/ official NNP/ Nintendo VBD/ announced NN/
today IN/ that PRP/ they MD/ Will VB/ release DT/ the NNP/ Nintendo NN/ 3DS IN/ in RB/
north NNP/ America NN/ march CD/ 27 IN/ for NN/ $250

Chunk:

[NP Yess] ! [NP Yess] ! [NP Its offcial Nintendo] [VP announ
[NP today] [SBAR that] [NP they] [VP Will release] [NP the Nin
[PP in] [NP north America march] [NP 27] [PP for] [NP $2

Noisy Style

NER:

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

NER: Rich Classes

Source: Strauss, Toma, Ritter, de Marneffe, Xu
Results of the WNUT16 Named Entity Recognition Shared Task (WNUT@COLING 2016)

NER: Genre Differences

	News	Tweets
PER	Politicians, business leaders, journalists, celebrities	Sportsmen, actors, TV personalities, celebrities, names of friends
LOC	Countries, cities, rivers, and other places related to current affairs	Restaurants, bars, local landmarks/areas, cities, rarely countries
ORG	Public and private companies, government organisations	Bands, internet companies, sports clubs

Source: Kalina Bontcheva and Leon Derczynski
"Tutorial on Natural Language Processing for Social Media" EACL 2014

Weakly Supervised NER

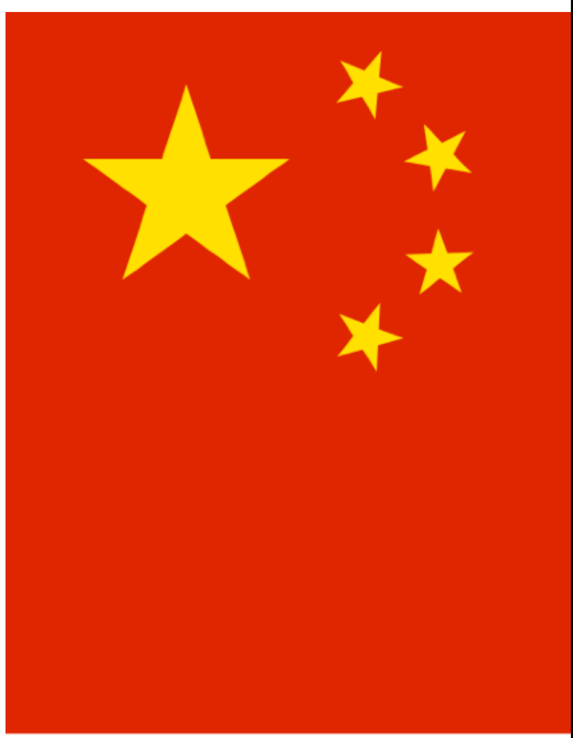
- Freebase / Wikipedia lists provide a source of supervision
- But these lists are highly ambiguous
- Example: **China**

Weakly Supervised NER

- Freebase / Wikipedia lists provide a source of supervision
- But these lists are highly ambiguous
- Example: **China**

Weakly Supervised NER

- Freebase / Wikipedia lists provide a source of supervision
- But these lists are highly ambiguous
- Example: **China**

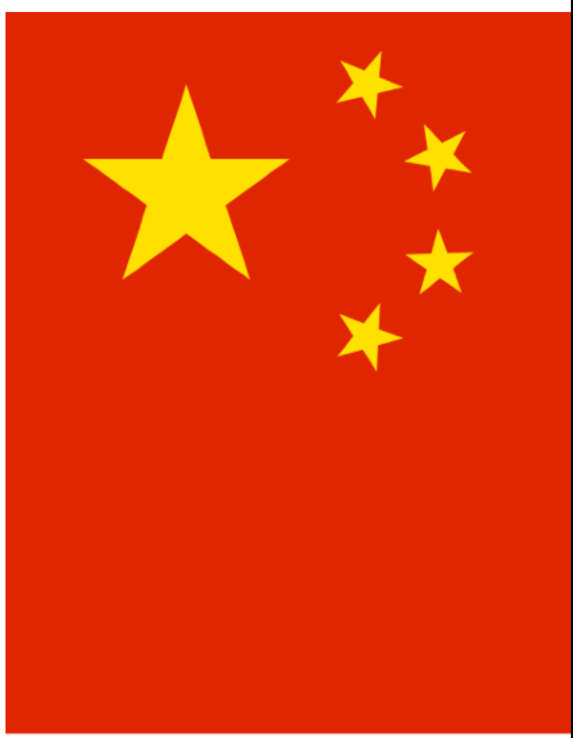


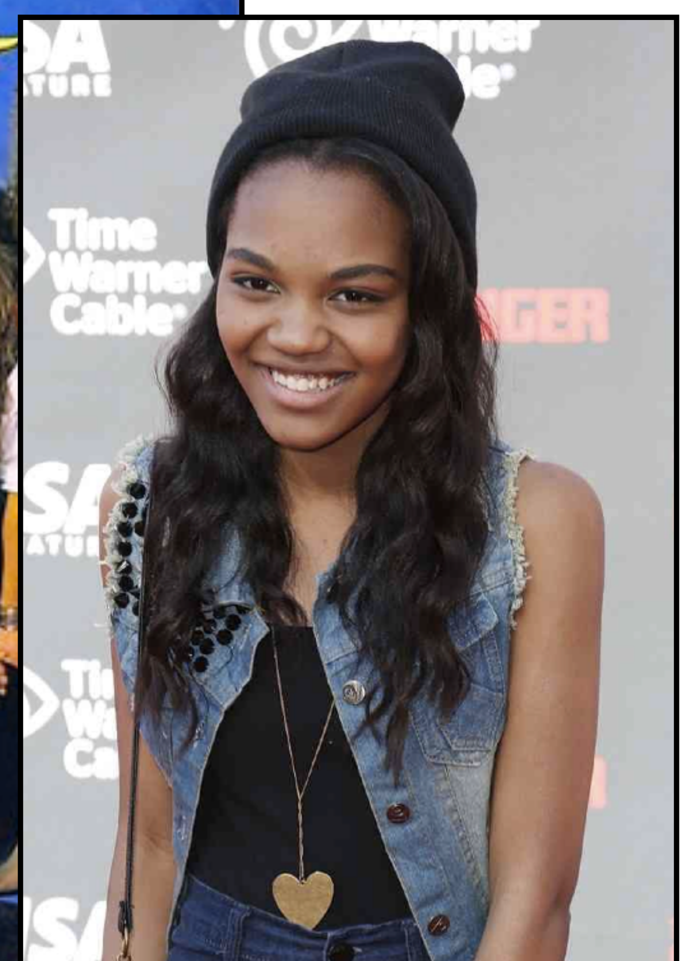
Weakly Supervised NER

- Freebase / Wikipedia lists provide a source of supervision
- But these lists are highly ambiguous
- Example: **China**

Freebase

Wikipedia
The Free Encyclopedia



Weakly Supervised NER

- Freebase / Wikipedia lists provide a source of supervision
- But these lists are highly ambiguous
- Example: **China**

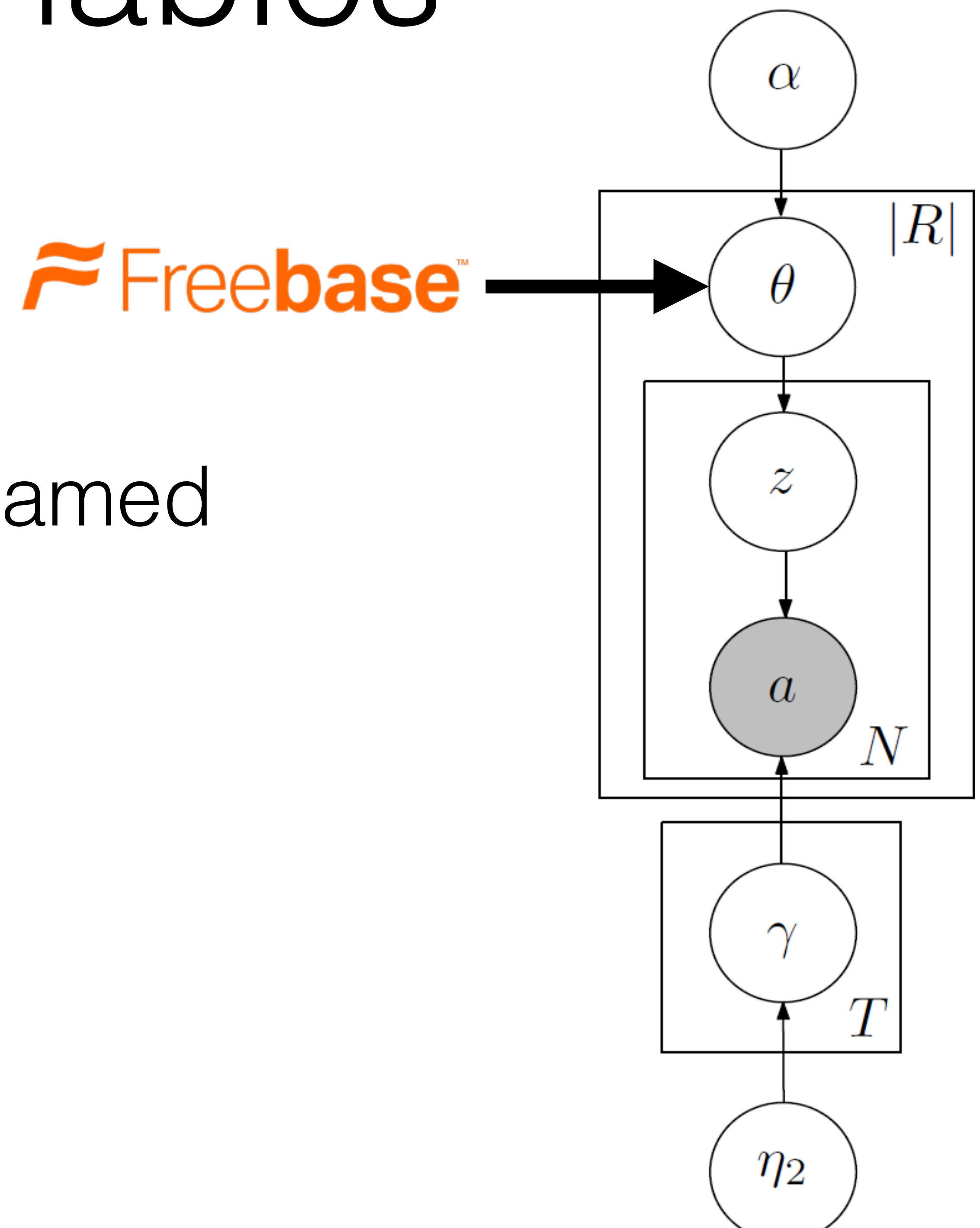
Freebase

Wikipedia
The Free Encyclopedia

…

[Ritter, et. al. EMNLP 2011]

Distant Supervision with Latent Variables



Latent variable model for Named Entity Categorization **with constraints**

[Ritter, et. al. EMNLP 2011]

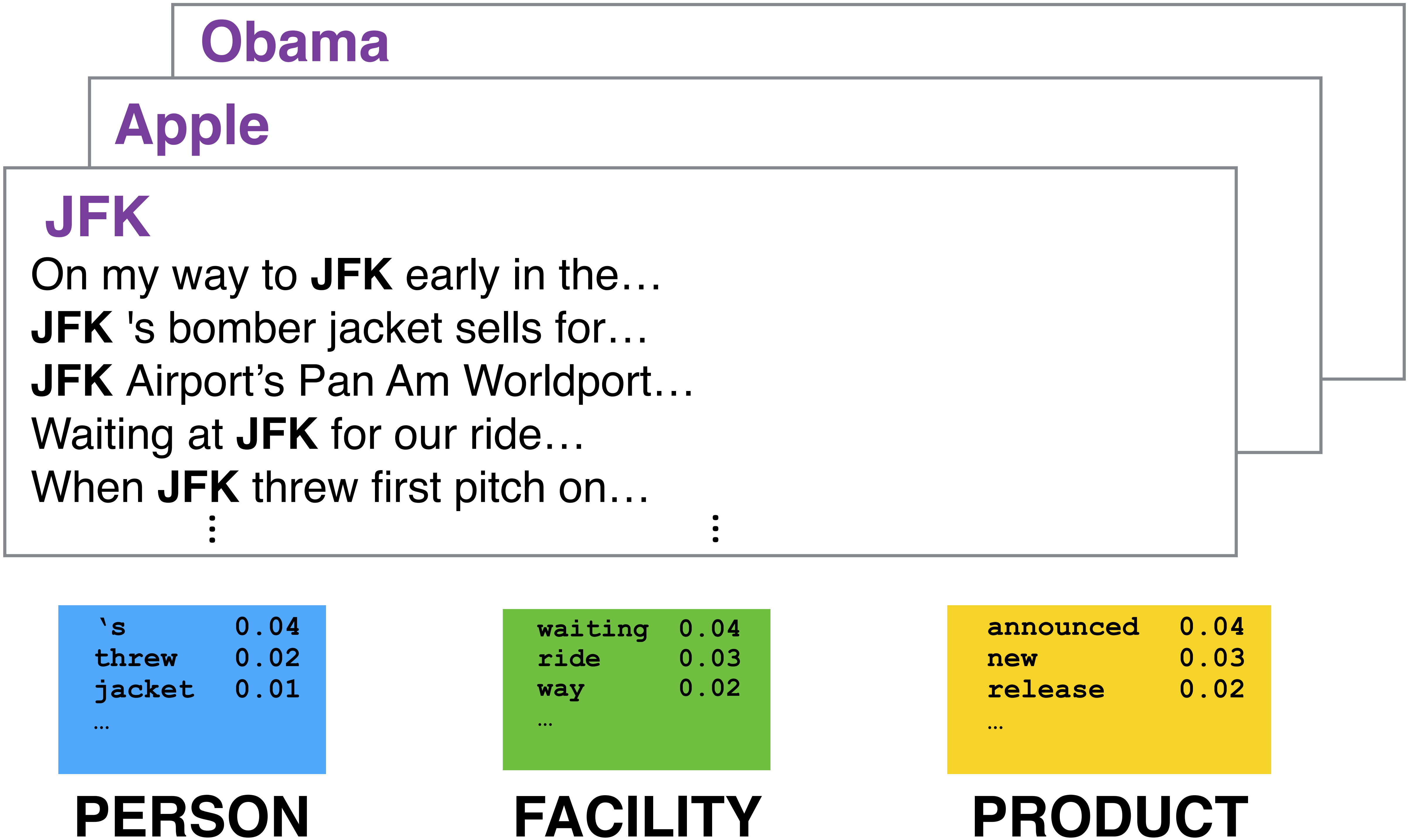
Obama

Apple

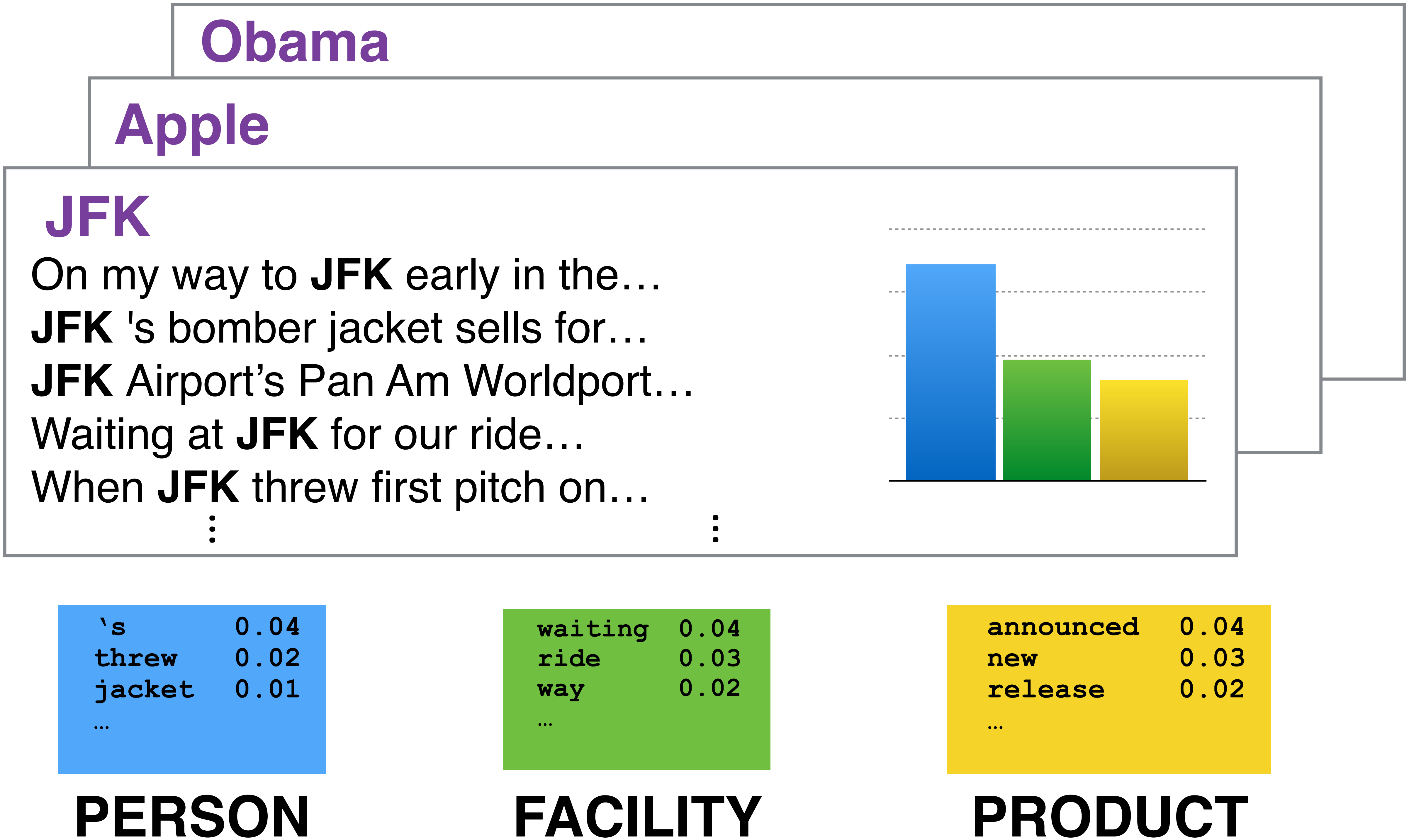
JFK

On my way to **JFK** early in the…
JFK 's bomber jacket sells for…
JFK Airport’s Pan Am Worldport…
Waiting at **JFK** for our ride…
When **JFK** threw first pitch on…
⋮ ⋮

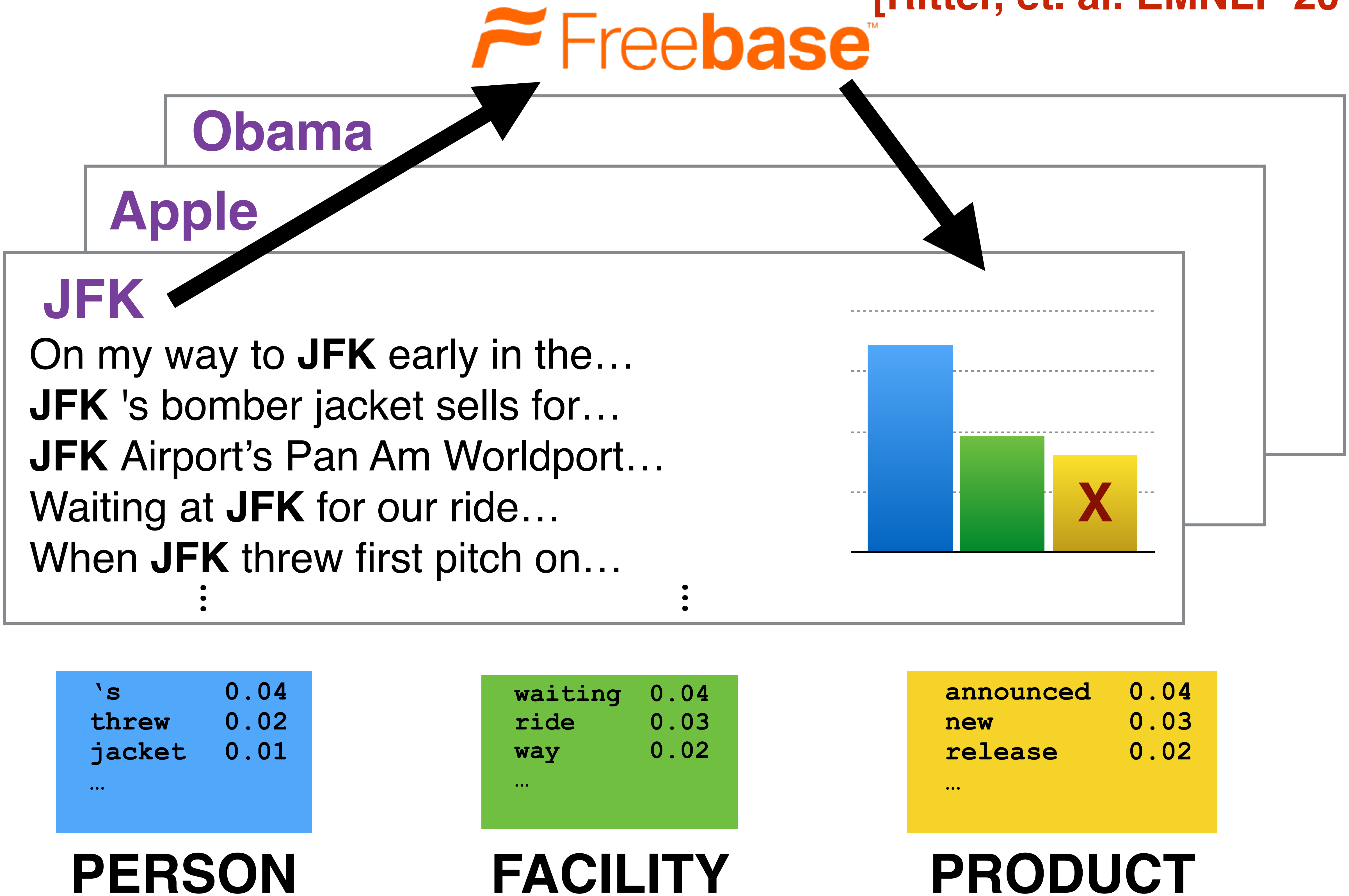
[Ritter, et. al. EMNLP 2011]



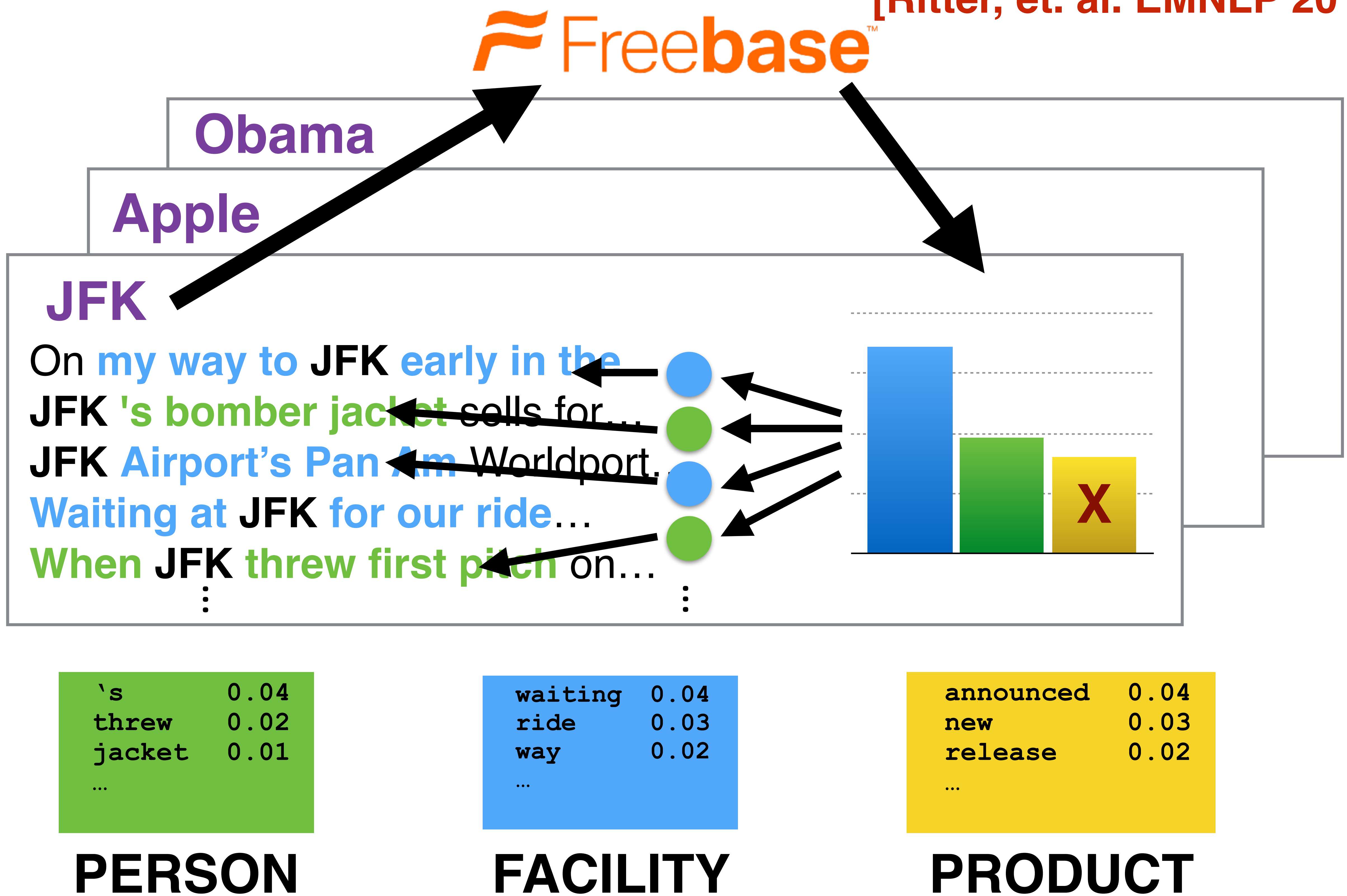
[Ritter, et. al. EMNLP 2011]



[Ritter, et. al. EMNLP 2011]



[Ritter, et. al. EMNLP 2011]



[Ritter, et. al. EMNLP 2011]

Example Type Lists

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

[Ritter, et. al. EMNLP 2011]

Example Type Lists

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, $\boxed{\text{kktny, rhobh,}}$ 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

KKTNY = Kourtney and Kim Take New York
RHOBH = Real Housewives of Beverly Hills

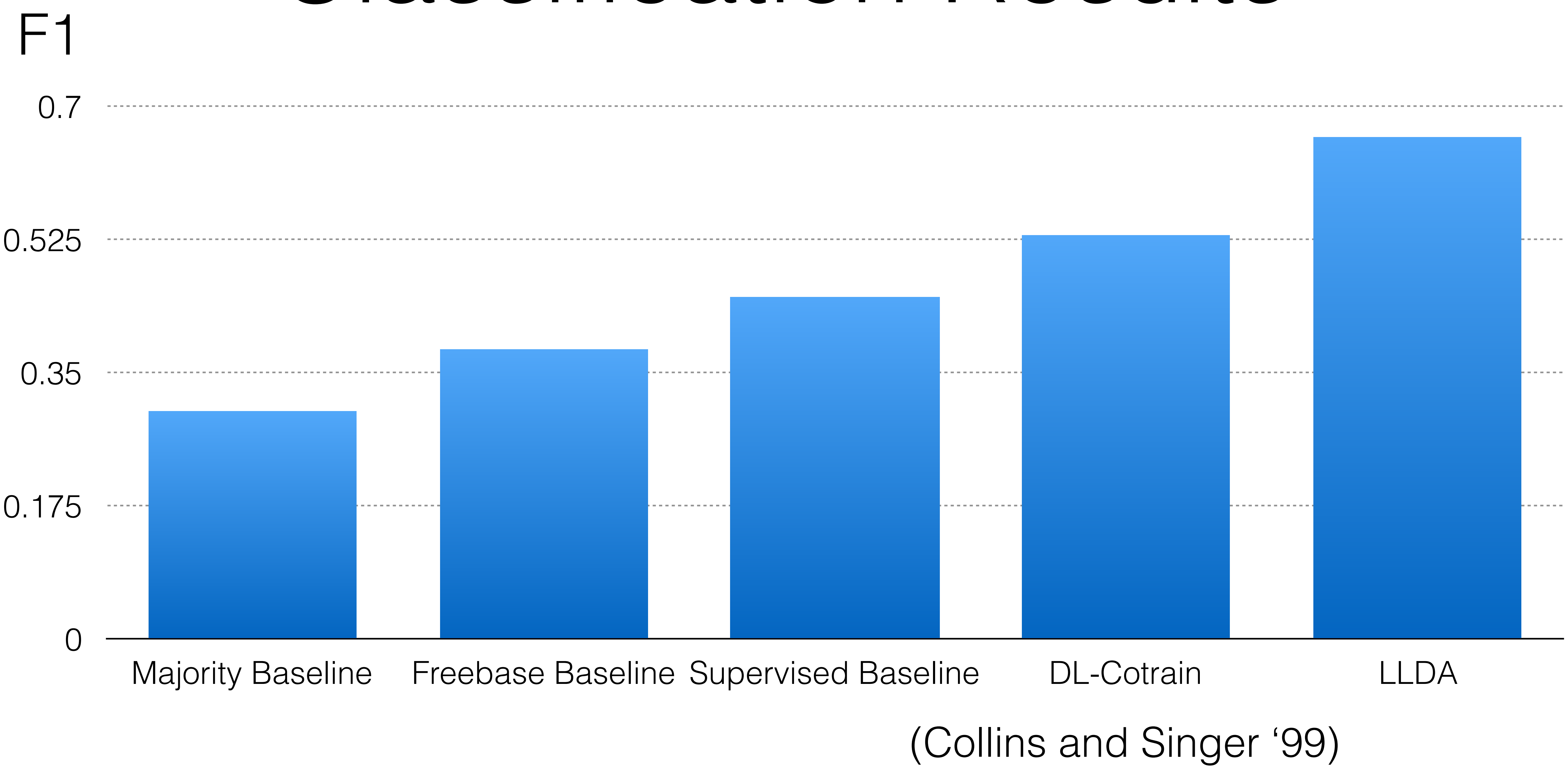
Example Type Lists

Type	Top 20 Entities not found in Freebase dictionaries
PRODUCT	nintendo ds n black, ipod nano, apple iphone, erizon media, mac app store, kde, h phone/ipod, galaxy tab, samsung gal nintendo ds, vpn
TV-SHOW	pretty little, svu, greys, kktny, rhobh, parks & recr n 's creek, big fat gypsy weddings, bi r wipeout, jersey shores, idiot abroad, royle, jerseyshore, mr . sunshine, hawaii five-0, new jersey shore
FACILITY	voodoo loun mansion, sullivan hall, memorial un l music hall, amway center, el moca gestone arena, cat club, le poisson ro y bay, broadway bar, ritz carlton, mgm grand, olympia theatre, consol energy center

KKTNY = Kourtney and Kim Take New York
RHOBH = Real Housewives of Beverly Hills

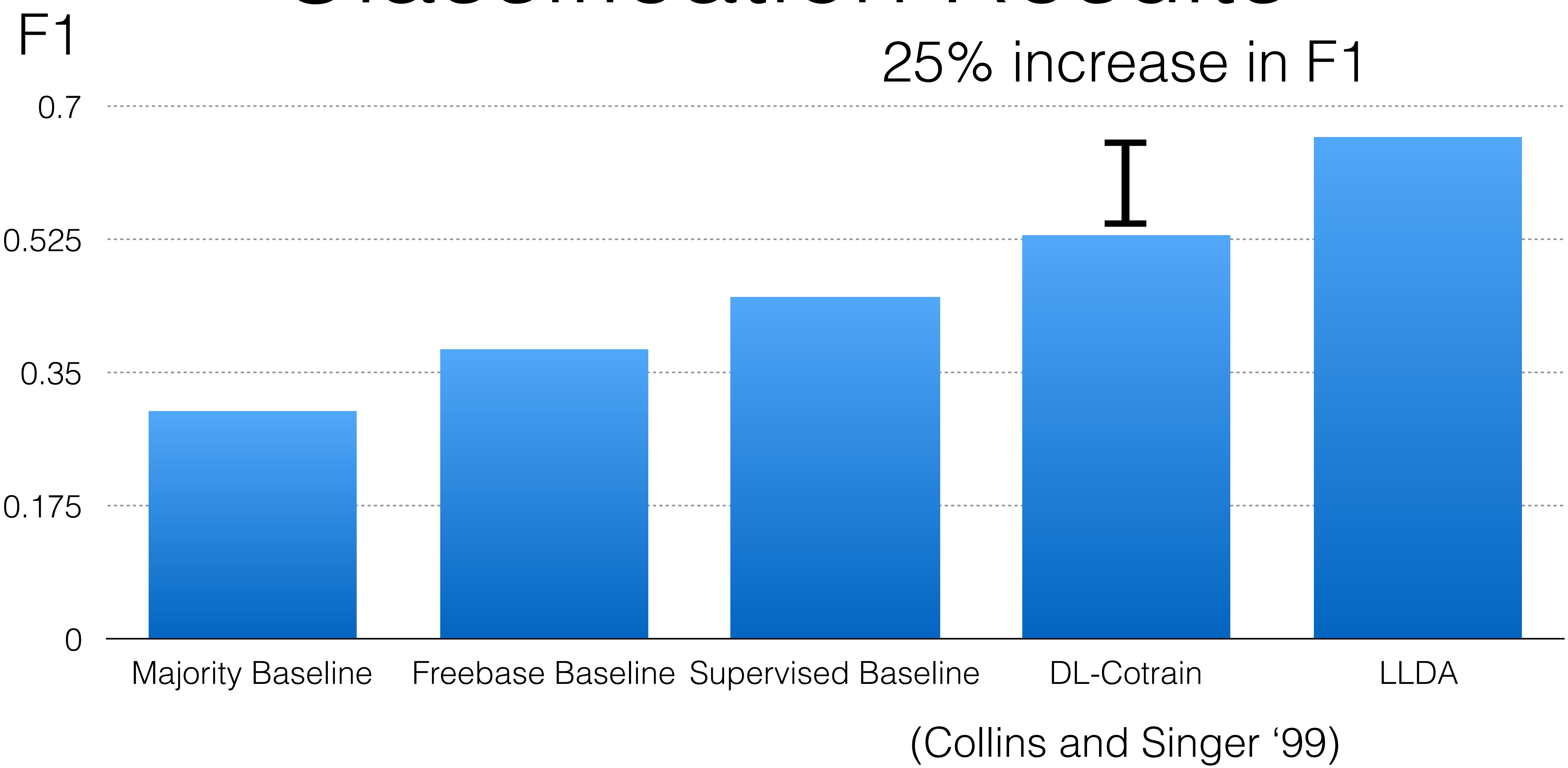
[Ritter, et. al. EMNLP 2011]

Twitter NER: Classification Results

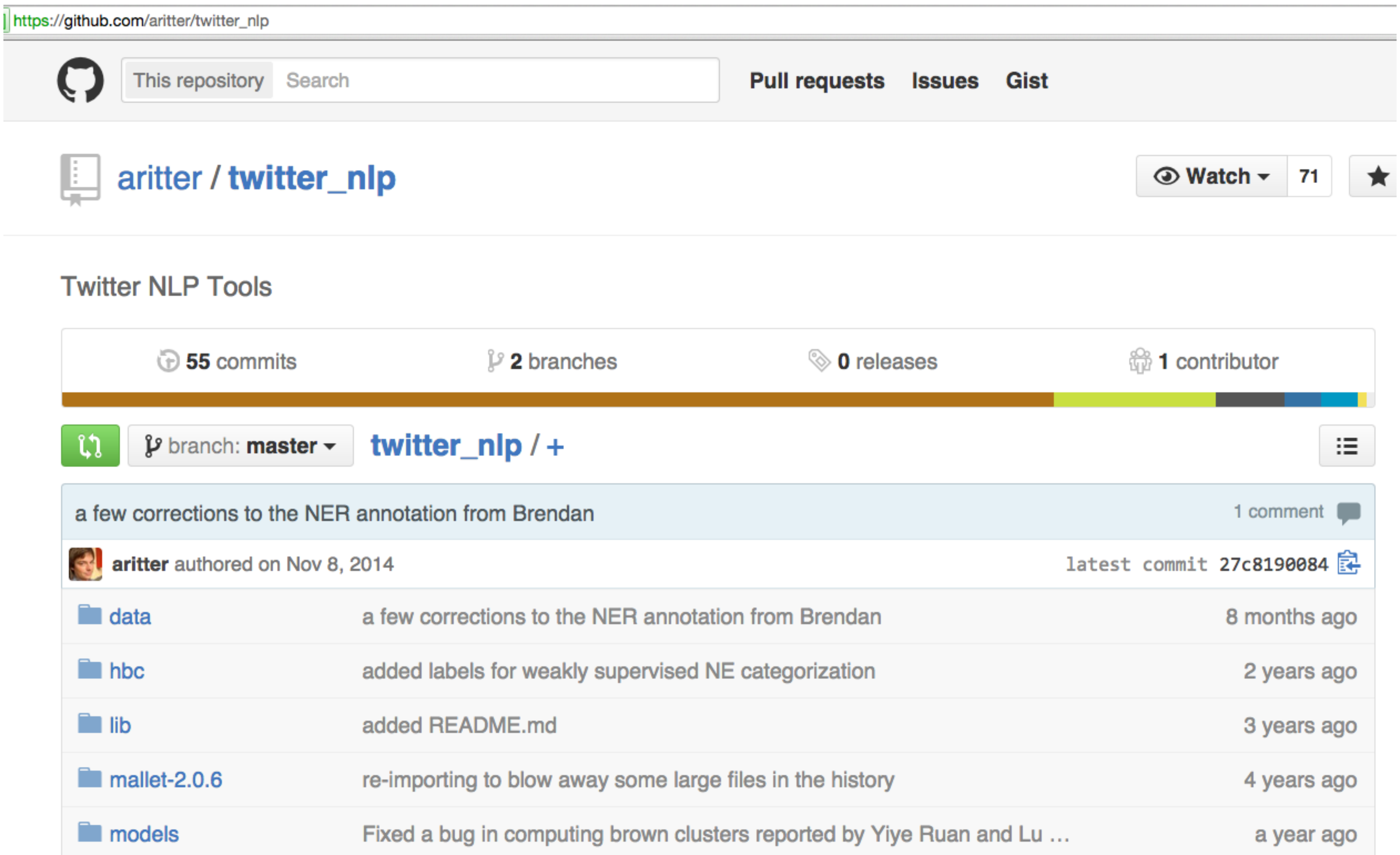


[Ritter, et. al. EMNLP 2011]

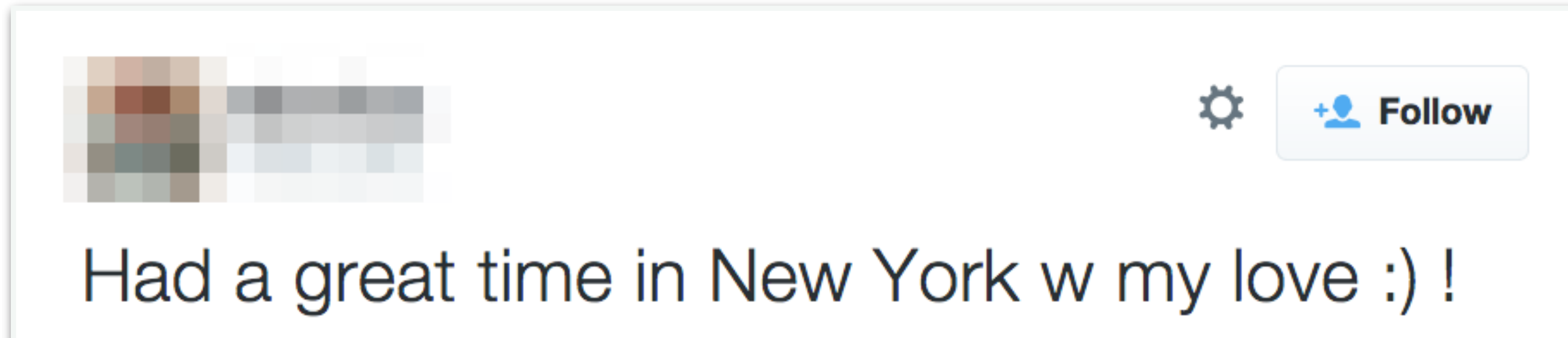
Twitter NER: Classification Results



Tool: twitter_nlp



Tool: twitter_nlp




```
xuwei@proteus100[twitter_nlp]$ export TWITTER_NLP=./
xuwei@proteus100[twitter_nlp]$
xuwei@proteus100[twitter_nlp]$ echo "Had a great time in New York w my
love :) ! " | python python/ner/extractEntities2.py

Had/O a/O great/O time/O in/O New/B-ENTITY York/I-ENTITY w/O my/O love/
O :)/O !/O
Average time per tweet = 3.04769945145s
xuwei@proteus100[twitter_nlp]$
xuwei@proteus100[twitter_nlp]$ echo "Had a great time in New York w my
love :) ! " | python python/ner/extractEntities2.py --pos --chunk

Had/O/VBD/B-VP a/O/DT/B-NP great/O/JJ/I-NP time/O/NN/I-NP in/O/IN/B-PP
New/B-ENTITY/NNP/B-NP York/I-ENTITY/NNP/I-NP w/O/IN/B-PP my/O/PRP$/B-NP
 love/O/NN/I-NP :)/O/UH/B-INTJ !/O/./I-INTJ
Average time per tweet = 5.49846148491s
xuwei@proteus100[twitter_nlp]$ _
```

Results of the WNUT16 Named Entity Recognition Shared Task

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine de Marneffe and Wei Xu

Need for Shared Evaluations

- **Fast Moving Area:** Papers published in the same year use different datasets and evaluation methodology
- Performance still behind what we would like
  - ~0.6 - 0.7 F1 score (**much lower than news**)
  - Explore new ideas & approaches

Related NER Evaluations

- MUC Newswire
  - http://www.itl.nist.gov/iaui/894.02/related_projects/muc/muc_data/muc_data_index.html
- CONLL Newswire
  - http://www.cnts.ua.ac.be/conll2002/ner/
  - http://www.cnts.ua.ac.be/conll2003/ner/
- ACE Newswire
  - https://catalog.ldc.upenn.edu/LDC2005T09
- Named Entity rEcognition and Linking (NEEL) Challenge Microblogs
  - #Microposts workshop at WWW
  - http://microposts2016.seas.upenn.edu/challenge.html

Twitter NER Evaluation Summary

Re-Run of 2015 Task

2 Subtasks

- Segmentation + 10 way classification
- Segmentation only (no classification)

Twitter NER Evaluation Summary

Re-Run of 2015 Task

2 Subtasks

- Segmentation + 10 way classification
- Segmentation only (no classification)

New test set annotated for 2016

Twitter NER Evaluation Summary

Re-Run of 2015 Task

2 Subtasks

- Segmentation + 10 way classification
- Segmentation only (no classification)

New test set annotated for 2016

10 Participating Teams

Data

Training + Dev Data:

- All training, dev, test data from 2015
- **Training:** 2,394 tweets, **Dev:** 1,420 tweets

Data

Training + Dev Data:

- All training, dev, test data from 2015
- **Training:** 2,394 tweets, **Dev:** 1,420 tweets

Test Data

- 3,856 tweets
- Later Time Period
  - No overlap in time period with Training/Dev data

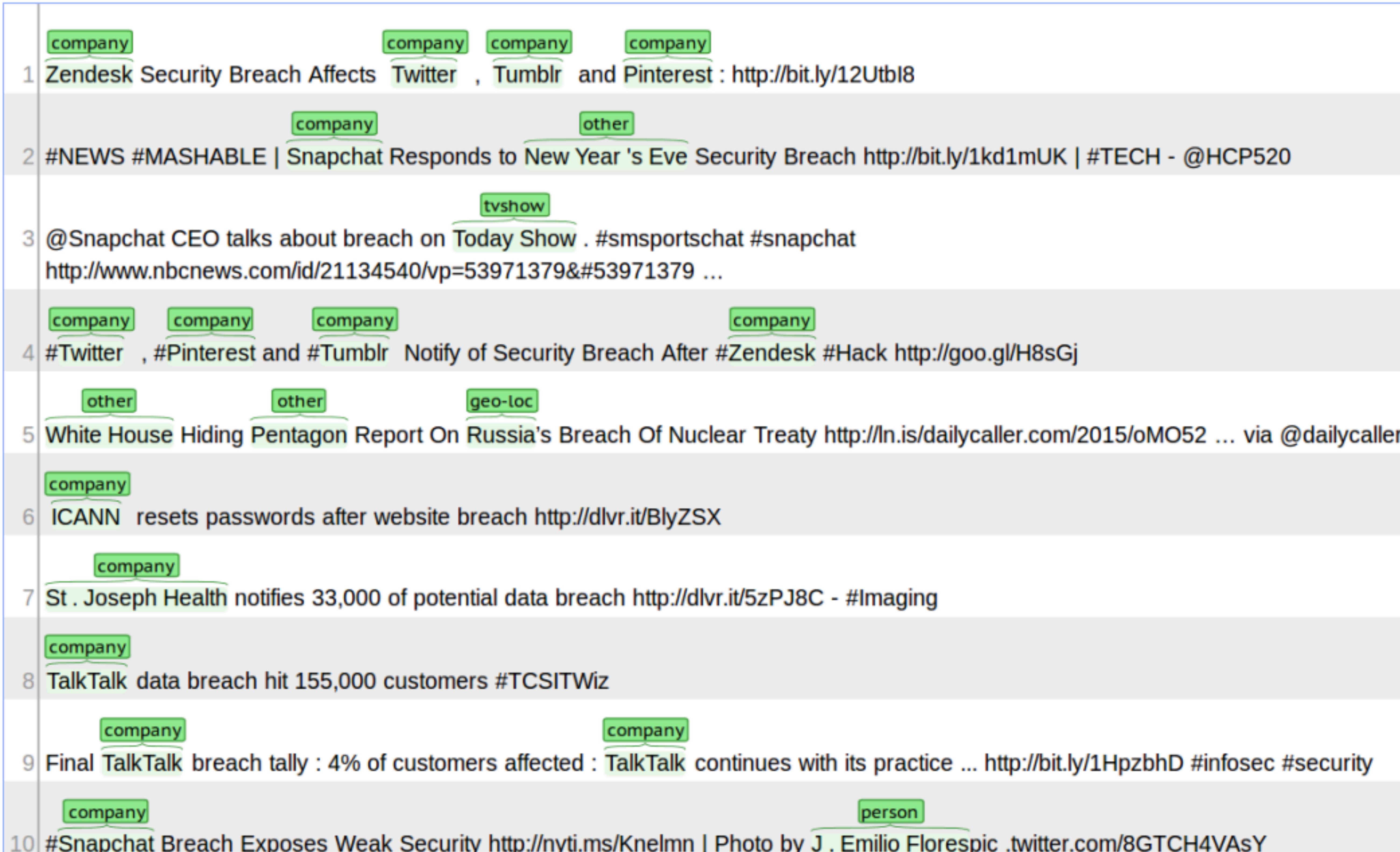
2016 Test Data Annotation

- Simple Annotation Guidelines:
  - http://bit.ly/1FSP6i2
- Re-annotated 100 tweets from 2015 data
  - Good agreement - 0.68 F-Score

2016 Test Data Annotation

- Simple Annotation Guidelines:
  - http://bit.ly/1FSP6i2
- Re-annotated 100 tweets from 2015 data
  - Good agreement - 0.68 F-Score
- Frequent questions, discussion among the group
- BRAT annotation tool
  - (http://brat.nlplab.org/)

Annotation Interface



10 Participating Teams

Team ID	Affiliation
CambridgeLTL	University of Cambridge
Talos	Viseo R&D
akora	University of Manchester
NTNU	Indian Institute of Technology Patna
ASU	Ain Shams University, Cairo, Egypt
DeepNNNER	Honda Research Institute Japan
DeepER	University of Illinois at Urbana-Champaign
hjpwhu	Wuhan University
UQAM-NTL	Université du Québec à Montréal
LIOX	The Hong Kong Polytechnic University

Approaches

	POS	Orthographic	Gazetteers	Brown clustering	Word embedding	ML
BASELINE	–	✓	✓	–	–	CRFsuite
CambridgeLTL	–	✓	–	–	–	LSTM
akora	–	–	–	–	–	LSTM
NTNU	✓	✓	✓	–	–	CRF
Talos	✓	✓	✓	✓	GloVe	L2S
DeepNNNER	–	–	–	–	Multiple	LSTM-CNN
ASU	–	–	✓	✓	–	LSTM
UQAM-NTL	✓	✓	✓	–	–	CRF

- All teams use some form of machine supervised ML
- Many LSTM-based approaches as compared to last year
- **Unique approaches:** CambridgeLTL, Talos

Results (10 types)

	Precision	Recall	F1
CambridgeLTL	60.77	46.07	52.41
Talos	58.51	38.12	46.16
akora	51.70	39.48	44.77
NTNU	53.19	32.13	40.06
ASU	40.58	37.58	39.02
DeepNNNER	54.97	28.16	37.24
DeepER	45.40	31.15	36.95
hjpwhu	48.90	28.76	36.22
UQAM-NTL	40.73	23.52	29.82
LIOX	40.15	12.69	19.26

Results (No Types)

	Precision	Recall	F1
CambridgeLTL	73.49	59.72	65.89
NTNU	64.18	62.28	63.22
Talos	70.53	52.58	60.24
akora	64.75	54.28	59.05
ASU	57.55	52.98	55.17
DeepER	63.17	43.31	51.38
DeepNNNER	70.66	36.14	47.82
hjpwhu	63.00	37.06	46.66
UQAM-NTL	53.21	37.95	44.30
LIOX	58.18	31.33	40.73

Domain-Specific Data

Cybersecurity (350 Tweets)

HACKMAGEDDON
Information Security Timelines and Statistics

Gun Violence (500 Tweets)

GUN VIOLENCE Archive

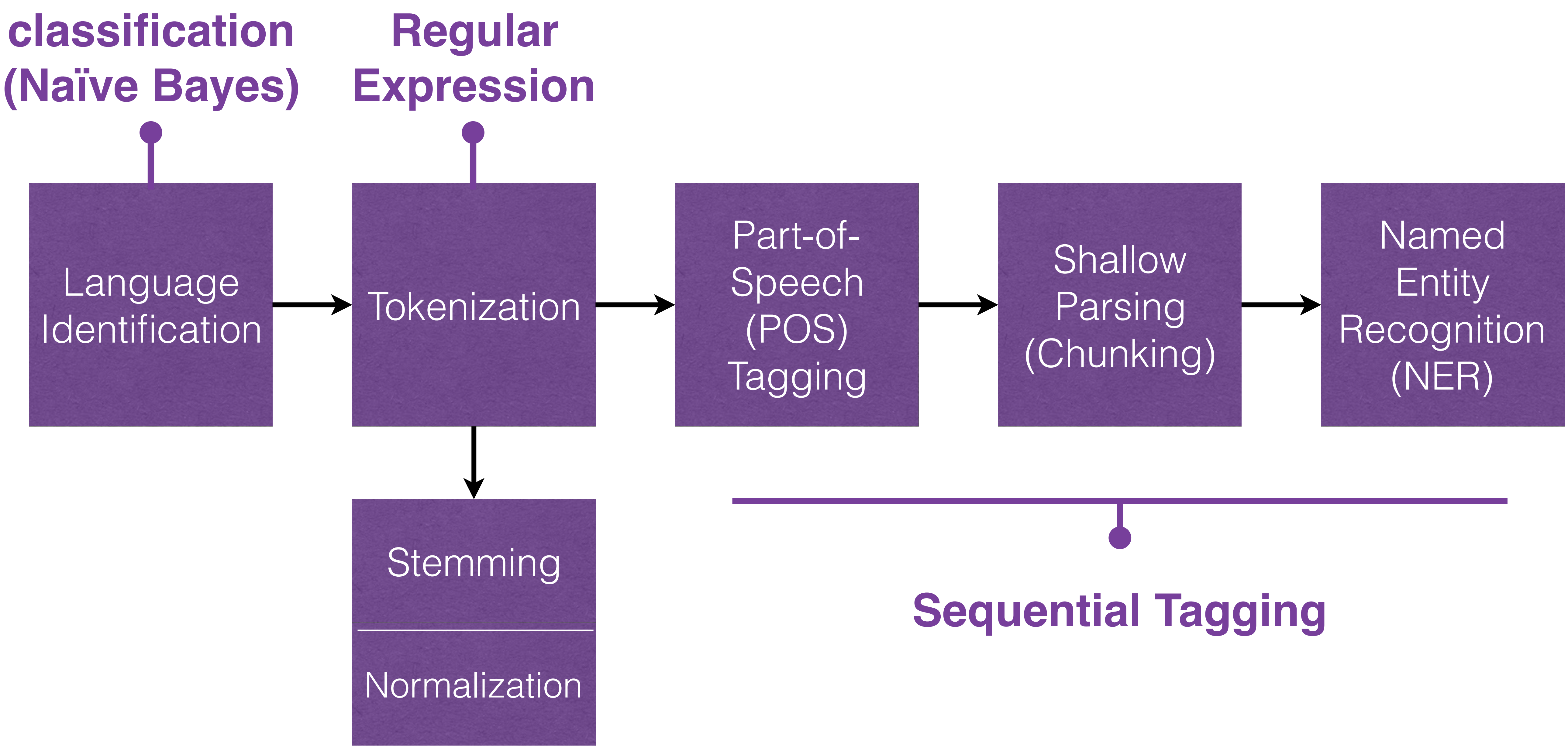
Results (Cyber-Domain)

	Acc	P	R	F1
CambridgeLTL	90.57	69.75	51.24	59.08
Talos	89.36	60.49	41.13	48.96
akora	88.42	54.21	36.32	43.50
hjpwhu	88.21	59.79	28.86	38.93
ASU	87.76	42.22	32.84	36.94
NTNU	87.72	51.89	27.36	35.83
DeepNNNER	87.66	62.88	23.88	34.62
DeepER	84.32	40.23	22.89	29.18
UQAM-NTL	85.64	37.97	16.75	23.25
LIOX	84.41	30.08	6.63	10.87

Results (Shooting-Domain)

	Acc	P	R	F1
CambridgeLTL	93.00	66.25	56.72	61.12
Talos	92.03	68.53	49.00	57.14
DeepER	91.96	64.01	51.40	57.02
akora	91.54	58.89	49.40	53.73
NTNU	91.14	61.36	42.08	49.92
DeepNNNER	91.22	59.88	41.15	48.78
hjpwhu	90.83	53.71	41.41	46.77
ASU	90.74	45.40	47.94	46.63
UQAM-NTL	89.38	45.80	33.42	38.65
LIOX	88.35	55.77	23.17	32.74

Summary


classification
(Naïve Bayes)
Regular
Expression
Language
Identification
Tokenization
Part-of-
Speech
(POS)
Tagging
Shallow
Parsing
(Chunking)
Named
Entity
Recognition
(NER)
Stemming
Normalization
Sequential Tagging

Presentation 2