

# GATE and Social Media: Named entities

Dom Rout  
Kalina Bontcheva  
Leon Derczynski (slides)

Texts frequently focus on particular entities

To discover what documents say about them, we can:

- Recognise entity mentions
- Disambiguate entities to external vocabularies
- Find opinions that authors have about the entities

Important:

- Enables IE over tweets
- Critical for event extraction (actors, events)
- Describes the topic of the tweet

Tough:

- ANNIE doesn't do well – around 50% F1
- Stanford's leading tool does even worse – around 40% F1!

What's going on? How can we build a tweet NER tool?

# NER Intro

---

- .We know social media is more diverse than canonical news text
- .We've browsed through entities in tweets
- .What practical issues are there in twitter NER?
- .What solutions have been proposed?

# Named entity recognition: example

Goal is to find mentions of entities

Newsire London Fashion Week grows up – but mustn't take itself too seriously. Once a launching pad for new designers, it is fast becoming the main event. But LFW mustn't let the luxury and money crush its sense of silliness.

---

Social media Gotta dress up for london fashion week and party in style!!!



# Named entity recognition: example

## Person mentions in news

Left context	Match	Right context
in dicated Atef, including	Douglas Feith	, the United States defence
, the group that killed	President Sadat	in 1981 as retribution for
. The current leader,	President Olusegun Obasanjo	, who recently came to
Kuwait, whose information minister	Sheikh Ahmed Fahed al-Sabah	met editors of local newspapers
The current defence minister,	Theophilus Danjuma	, has also been threatened
The three right-wing MPs,	Andrew Rosindell	(Romford), Andrew
Late on Wednesday night,	Justice Oputa	, who chairs the commission
the militarily-manoevred civilian elec...	President Obasanjo	in 1999 and is widely
after the mysterious death of	General Sani Abacha	in 1998.
have learnt that one of	Bin Laden	's closest and most senior
evidence confirms the involvement of	Osama bin Laden	in those attacks."
. He is one of	Bin Laden	's two most senior associates
for future civilian office.	General Buhari	took power in a 1983
\$5m price on	Atef	's head and prosecutors have
Afghanistan. He was once	Bin Laden	's chief media adviser and
thinking in the Tory party	Iain Duncan Smith	has ordered three Tory MPs
club and the party,	David Maclean	, the Tory Chief Whip
Centre and the Pentagon.	Mohammed Atef	, who is thought to
are still very powerful.	General Babangida	supported the militarily-manoevred ci
sexual orientation or religion.	Mr Duncan Smith	's purge of the Monday
, " he said.	Atef	, who is reported variously
of the late singer,	Fela Kuti	✎ which took place while
field in Penn sylvania.	President Bush	included Atef in an order
. It is believed that	Mr Duncan Smith	intended to launch his crackdown

# Named entity recognition: example

## Person mentions in tweets

Left context	Match	Right context
i was your age ,	spencer	from iCarly was Crazy Steve
iCarly was Crazy Steve ,	Carly	was Megan and Josh was
bath , shut up ,	sam	's coming tomorrow and steve
. All are welcome ,	joe	included
. All are welcome ,	joe	included
teachers , chinese takeaways ,	gatt holly	, phil collins , the
takeaways , gatt holly ,	phil collins	, the skin of a
@GdnPolitics : RT AlJahom :	Blair	: " I'm gonna
Empls of the Month :	Deborah L	#Speech #Pathologist-Childrens
be the next Pope "	Brown	: " I won't
( via POPSUGAR )	Sarah Jessica Parker	and Gwen Stefani Wrap Up
and is smexy !!; )And	Chelsea Handler	is hilarious ! Finally got
him befnrjustthen about	kenny	signing his book but it
three kinds of reactions after	Ayodhya	verdict .
, Carly was Megan and	Josh	was fat . #damnteenquotes
sam 's coming tomorrow and	steve	and tanya will be round
coming tomorrow and steve and	tanya	will be round at 10am
photo caption contest- Nadal and	Novak	in the tub <a href="http://ow.ly/2G3Jh">http://ow.ly/2G3Jh</a>
) Sarah Jessica Parker and	Gwen Stefani	Wrap Up Another Successful New
#Pathologist-Childrens Rehab and	Patricia M	#Referral/#Auth #
Just casually stalking Cheryl AND	Dermot	tomorrow .... NO BIGGIE
did tweet him befnr	justthen	about kenny signing his book
Test : We just congratulated	Lindsay	an hour ago on h
the funnv photo caption contest-	Nadal	and Novak in the tub

# Named entity recognition: resources

---

## UW (Ritter, 2011)

- 34k tokens, 1500 entities
- Single annotator
- Ten entity types: PERSON, GEO-LOCATION, COMPANY, PRODUCT, FACILITY, TV-SHOW, MOVIE, SPORTSTEAM, BAND, and OTHER

## UMBC (Finin, 2010)

- 7k tokens, 500 entities
- Multiple annotator
- Three entity types: PERSON, LOCATION, ORGANISATION

## MSM2013 (Basave, 2013)

- 30k tokens, 1500 entities
- Multiple annotator
- Three entity types: PERSON, LOCATION, ORGANISATION
- Hashtags, URLs and entities obfuscated

# Named entity recognition: issues

Genre differences in entity type

	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



# Named Entity Recognition Structure

---

Design choices in NER: (Roth 2009)

- **What feature representation to use for tokens;**
- **Which inference algorithm to use;**
- **How to capture non-local dependencies;**
- **How to incorporate external knowledge.**

# Representation and labeling

---

Token feature representation options:

- Token itself
- Previous and following token
- Word shape, to model capitalisation
- Lexical features (e.g. character n-grams) to help with OOV terms
- Part of speech tag
- Parsing information
- Word embeddings
- Word sense clusters

NER inference algorithms

As with part of speech tagging, sequence labelling can work well (e.g. CRF)

- Assumes well-formed sentences and lots of training data
- If this is inappropriate, then local context in token features can compensate

# Representation and labeling

Labelling scheme:

I	Facebook	B-company
O	Job-Hunting	O
O	App	O
I	BranchOut	B-product
O	Raises	O
O	\$6	O
O	Million	O
O	From	O
I	Accel	B-company
O	And	O
I	Super	B-company
I	Angels	I-company

BIO (Begin, In, Out) allows separation of adjacent entities

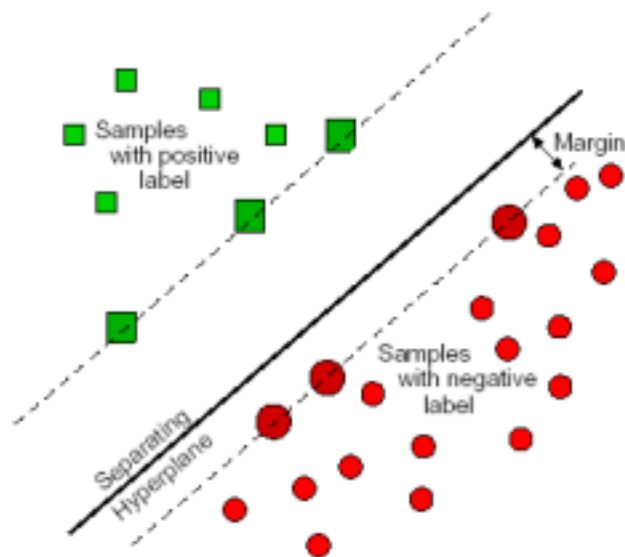
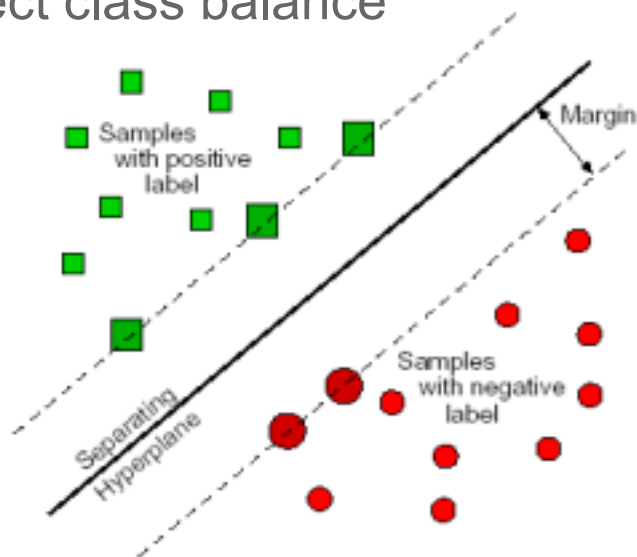
CRF with BIO popular

SVM-U with IO can give better performance

# Representation and labeling

## SVM-U: “uneven” (Li 2009)

Adjust margins between supporting examples and decision hyperplane to reflect class balance



Well-suited to tasks like NER, where one class is much more frequent than another

Retains SVM's advantage of being noise-resistant

# Dependencies & external knowledge

Typically, only the first mention of an entity is referred to in full:

Manchester United are great. They're my favourite football team. Man U  
forever!

Using only local features will lead to missed entities.

Tweets are not long discourses

- Possible for the long first mention to be missing
- Include context from elsewhere

How can we incorporate external knowledge for NER?

- Useful for unusual/unexpected words in an entity: “Szeged” “White House”

# Dependencies & external knowledge

---

## Unlabelled text

- NEs found in distributionally similar contexts
- Labelled LDA can produce phrase lists given an entity type (Ramage 2009, Ritter 2011)

## Gazetteers

- Can be constructed manually or automatically
- Gaz. completeness gives P/R tradeoff
- Won't catch terms not seen in gazetteer, which makes domain adaptation tough

# Named entity recognition approaches

---

Ritter (2011) addresses named entity recognition in tweets using a data-intensive approach

Distinct segmentation and classification tasks

- Discriminative segmentation
- Distantly supervised classification

Assume that @mentions are unambiguous

Found that inclusion out-of-domain data (from MUC) actually reduces performance

# Named entity recognition approaches

Models entity segmentation as sequence labeling using BIO representation and CRF

- Orthographic, contextual features
- Dictionary features based on type lists in Freebase
- Brown clusters from PoS tagging, NP/VP/PP chunking, capitalisation

Segmentation outperforms default Stanford NER consistently

- Stanford: F1 44%
- Segmentation without clusters: F1 63%
- Segmentation with clusters: F1 67% (52% error reduction)



# Named entity recognition approaches

---

After segmentation, Ritter (2011) describes NE classification

- Diversity in entity types exacerbates data sparsity problem
- Lack of context makes classification difficult even for humans
- e.g., [KKTNY in 45min.....](#)
- Co-occurrence can help in situations like this (Downey 2010)

# Named entity recognition approaches

Exploiting co-occurrence information with LabeledLDA and Freebase

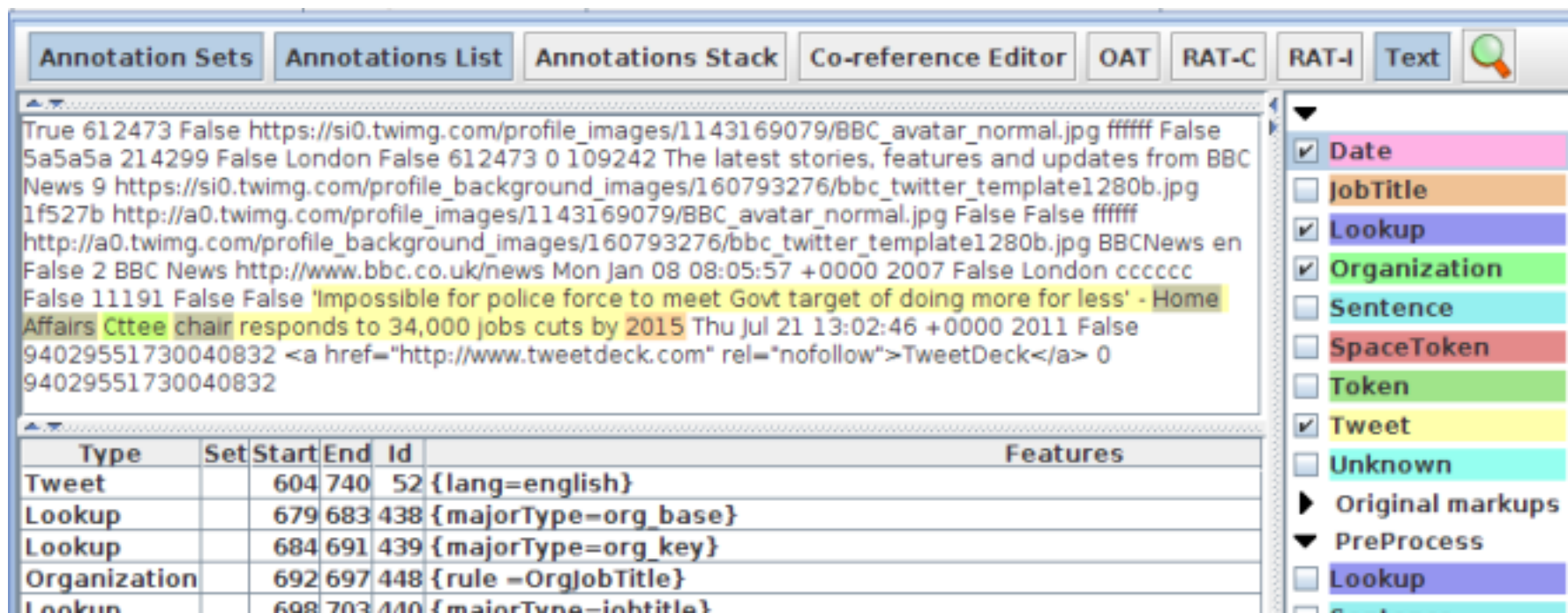
- .Freebase provides type ontology
- .LabeledLDA assigns distribution of potential Freebase types to entity mentions
- .Entity mention context modelled as bag-of-words
- .Distribution can vary from mention to mention
- .Include prior for type distribution  $\theta_e$  from encountered examples, to compensate for cases where there are few words for context

Evaluation over 2400 tweets, 10 types

- .Unlabelled data from 60M NE segmented tweets (24K distinct entity strings)
- .Freebase F1 38%
- .Supervised F1 45% (MaxEnt)
- .LabeledLDA F1 66%

# ANNIE NER on Tweets

- To run the ANNIE Transducer just on the tweet text:
  - Create and add an ANNIE Gazetteer and add it
  - Instantiate an ANNIE NE Transducer PR with defaults
  - Add it to the end of your application
  - Run it and inspect the default annotation set for NEs



The screenshot shows the GATE interface with the following components:

- Annotation Sets:** A list of annotation sets including Date, JobTitle, Lookup, Organization, Sentence, SpaceToken, Token, Tweet, and Unknown.
- Annotations List:** A table showing the results of the ANNIE NER application. The table has columns for Type, Set, Start, End, Id, and Features.
- Annotations Stack:** A list of annotations, including Tweet, Lookup, Organization, and Sentence.
- Co-reference Editor:** A tab for managing co-references.
- OAT:** A tab for the Open Annotation Tool.
- RAT-C:** A tab for the Rule Annotation Tool.
- RAT-I:** A tab for the Rule Annotation Tool.
- Text:** A tab for the text input.

The main window displays a tweet text with various entities highlighted in different colors. The right sidebar shows the 'Annotations List' tab with a list of annotation types and their corresponding features.

Type	Set	Start	End	Id	Features
Tweet		604	740	52	{lang=english}
Lookup		679	683	438	{majorType=org_base}
Lookup		684	691	439	{majorType=org_key}
Organization		692	697	448	{rule =OrgJobTitle}
Lookup		698	703	440	{majorType=jobtitle}

# Why the mistake? OrgJobTitle rule

Rule: OrgJobtitle

Priority: 30

```
(
  {Unknown.kind == PN} //It is only considering one preceding word as a candidate
                        //Grammar in plugins/ANNIE/resources/NE/org_context.jape
):org
(
  {Lookup.majorType == jobtitle}
)
-->
{
  gate.AnnotationSet org = (gate.AnnotationSet) bindings.get("org");
  gate.FeatureMap features = Factory.newFeatureMap();
  features.put("rule ", "OrgJobTitle");
  outputAS.add(org.firstChild(), org.lastNode(), "Organization",
               features);
  outputAS.removeAll(org);
}
```

# Tweet Capitalisation: an NER nightmare!

True 972651 True https://si0.twimg.com/profile\_images/58439629/petepassport\_normal.PNG 88dbf4  
 False 3b3b3b 2408043 False NYC / SF False 972651 -28800 37706 Breaking social media, tech and  
 digital news and analysis from Mashable.com, the top resource and guide for all things web. Updates  
 from @mashable staff. 2269  
 https://si0.twimg.com/profile\_background\_images/208575865/mashable\_main\_twitter\_bk\_v3.png 0f78c2  
 http://a1.twimg.com/profile\_images/58439629/petepassport\_normal.PNG False False fffef0  
 http://a2.twimg.com/profile\_background\_images/208575865/mashable\_main\_twitter\_bk\_v3.png mashable  
 en False 0 Pete Cashmore http://mashable.com Mon Mar 12 01:28:01 +0000 2007 False Pacific Time (US  
 & Canada) 5ea7db False 78468 False False Nokia Posts Huge Quarterly Loss, Sees Better Times Ahead -  
 http://on.mash.to/nCSH4i Thu Jul 21 13:12:30 +0000 2011 False 59 83 http://on.mash.to/nCSH4i  
 94031999962071040 <a href="http://www.hootsuite.com" rel="nofollow">HootSuite</a> 0  
 94031999962071040

Type	Set	Start	End	Id	Features
Organization		736	741	508	{orgType=company, rule1=GazOrganization, rule2=OrgFinal}
Organization		769	786	509	{rule1=TheOrgXKey, rule2=OrgFinal}

- ☐ Lookup
- ☒ Organization
- ☐ Sentence
- ☐ SpaceToken
- ☐ Token
- ☐ Tweet
- ☐ URL
- ☐ Unknown
- Original markups
- ▼ PreProcess
  - ☐ Lookup
  - ☐ Sentence

#WiredBizCon #nike vp said when @Apple saw what http://nikeplus.com did, #SteveJobs was  
 like wow I didn't expect this at all

...And hashtag semantics is yet another...

## Case-Insensitive matching

---

- This would seem the ideal solution, especially for gazetteer lookup, when people don't use case information as expected
- However, setting all PRs to be case-insensitive can have undesired consequences
  - POS tagging becomes unreliable (e.g. “May” vs “may”)
  - Back-off strategies may fail, e.g. unknown words beginning with a capital letter are normally assumed to be proper nouns
  - BUT this doesn't work on tweets anyway!
  - Gazetteer entries quickly become ambiguous (e.g. many place names and first names are ambiguous with common words)
- Solutions include selective use of case insensitivity, removal of ambiguous terms from lists, additional verification (e.g. use of the text of any contained URLs)









## More flexible matching techniques

---

- In GATE, as well as the standard gazetteers, we have options for modified versions which allow for more flexible matching
- BWP Gazetteer: uses Levenshtein edit distance for approximate string matching
- Extended Gazetteer: has a number of parameters for matching prefixes, suffixes, initial capitalisation and so on

## Try: Run ANNIE on User Profile Text

- User descriptions are another piece of useful text to mine
- Appear as UserDescription annotations in PreProcess
- Create another Annotation Set Transfer from PreProcess to the default set, using the UserDescription annotation from PreProcess as the textTagName
- HINT: See the parameters of the Tweet POS AST
- Add the new AST PR after the Tweet POS AST, but before the TwitIE POS Tagger. Re-run the app

	 Tweet POS AST	Annotation Set
	 UserDescr AST	Annotation Set
	 TwitIE POS Tagger	Stanford Tagge
	 NE ANNIE NE Transducer_000A8	ANNIE NE Trans



# ANNIE Results in User Descriptions

True 612473 False https://si0.twimg.com/profile\_images/1143169079/BBC\_avatar\_normal.jpg ffffff  
 False 5a5a5a 214299 False London False 612473 0 109242 The latest stories, features and  
 updates from BBC News 9  
 https://si0.twimg.com/profile\_background\_images/160793276/bbc\_twitter\_template1280b.jpg  
 1f527b http://a0.twimg.com/profile\_images/1143169079/BBC\_avatar\_normal.jpg False False ffffff  
 http://a0.twimg.com/profile\_background\_images/160793276/bbc\_twitter\_template1280b.jpg  
 BBCNews en False 2 BBC News http://www.bbc.co.uk/news Mon Jan 08 08:05:57 +0000 2007

True 20446311 False https://si0.twimg.com/profile\_images/662225518/professional\_normal.png F3F3F3  
 False 333333 1239 False Blacksburg, VA False 20446311 -21600 2404 Founder/CEO of Digital Bazaar.  
 RDFa/RDF WebApps Chair @ W3C. Champion for art/science, distributed banking/commerce, @PaySwarm,  
 JSON-LD, semantics and puppies. 160 https://si0.twimg.com/images/themes/theme7/bg.gif 990000  
 http://a0.twimg.com/profile\_images/662225518/professional\_normal.png False False EBEBEB  
 http://a1.twimg.com/images/themes/theme7/bg.gif manusporny en False 0 Manu Sporny  
 http://digitalbazaar.com/blog/ Mon Feb 09 16:33:16 +0000 2009 False Central Time (US & Canada)  
 DFDFDF False 189 False False Automotive RDFa (a horribly researched SEO article on RDFa/Microformats):  
 http://ow.ly/5JSoS #somanerrorsitsfunny Thu Jul 21 13:01:21 +0000 2011 False 93 114  
 somanerrorsitsfunny 74 92 http://ow.ly/5JSoS 94029193863639040 <a href="http://www.hootsuite.com"  
 rel="nofollow">HootSuite</a> 0 94029193863639040

Type	Set	Start	End	Id	Features
UserDescription		164	323	3297	{ rule=Description }
Person		222	233	3702	{ gender=[null], rule=PersonFinal, rule1=PersonFull }

☐ Date  
☐ JobTitle  
☐ Lookup  
☒ Organization  
☐ Hashtag  
☐ JobTitle  
☐ Lookup  
☒ Person  
☐ Sentence  
☐ SpaceToken  
☐ Title  
☐ Token  
☐ Tweet  
☐ URL  
☐ Unknown  
☒ UserDescription  
☐ UserID  
☐ Original markups  
☐ PreProcess

...TwitIE NE rules are being improved, watch this space...

# Hands-on: NER

Let's measure ANNIE performance on social media text

We'll run this over the Ritter-dev corpus, from r-tweets, so if you don't have this open, you can open it from the datastore saved in corpora/r-tweets

Run your pipeline, including the ANNIE NE transducers, on this corpus

Open the corpus and click the "Corpus Quality Assurance" tab

We want to compare Original Markups, the key, with the default annotations, the response

Select annotation types of Location, Organization, and Person

Pick an evaluation measure

How does it do? What kinds of errors are most prevalent, missed or spurious?

You can also pick individual documents and see which single annotations are picked up or missed

# Named entity recognition summary

---

Named entity recognition in tweets is hard

Three major classes of Tweet NER approach:

Sequence labelling – like Stanford CRF chunker

**Problem:** tweets aren't well-formed enough

**Problem:** lack of training data

Lookup-based using local grammar and string matching

**Problem:** strings are often misspelled

**Problem:** entity mentions aren't in gazetteers (drift) (Eisenstein 2013, Plank 2014)

**Advantage:** cuts through linguistic noise, agnostic to many style variations

Grounding to vocabulary (e.g. Dbpedia)

**Problem:** insufficient context to disambiguate

**Problem:** entities often appear in social media before the resource

# Overall solutions to twitter noise

---

## Normalisation

- Convert twitter text to “well-formed” text; e.g. slang resolution
- Some success using noisy channel model (Han 2011)
- Techniques include: edit distance; double metaphone with threshold
- Issues: false positives can change meanings, e.g. reversing sentiment (apolitical)

## Domain adaptation

- Treat twitter as its own genre, and create customised tools and techniques
- Some success in language ID (Carter 2013), PoS tagging (Gimpel 2011), NER (Ritter 2011)

## Flexible representations

- Represent document text in a way that is resilient to noise
- Brown clustering, word embeddings, deep learning