Part 2: Social Media Analysis: Problems and Solutions

Gartner 3V definition of Big Data

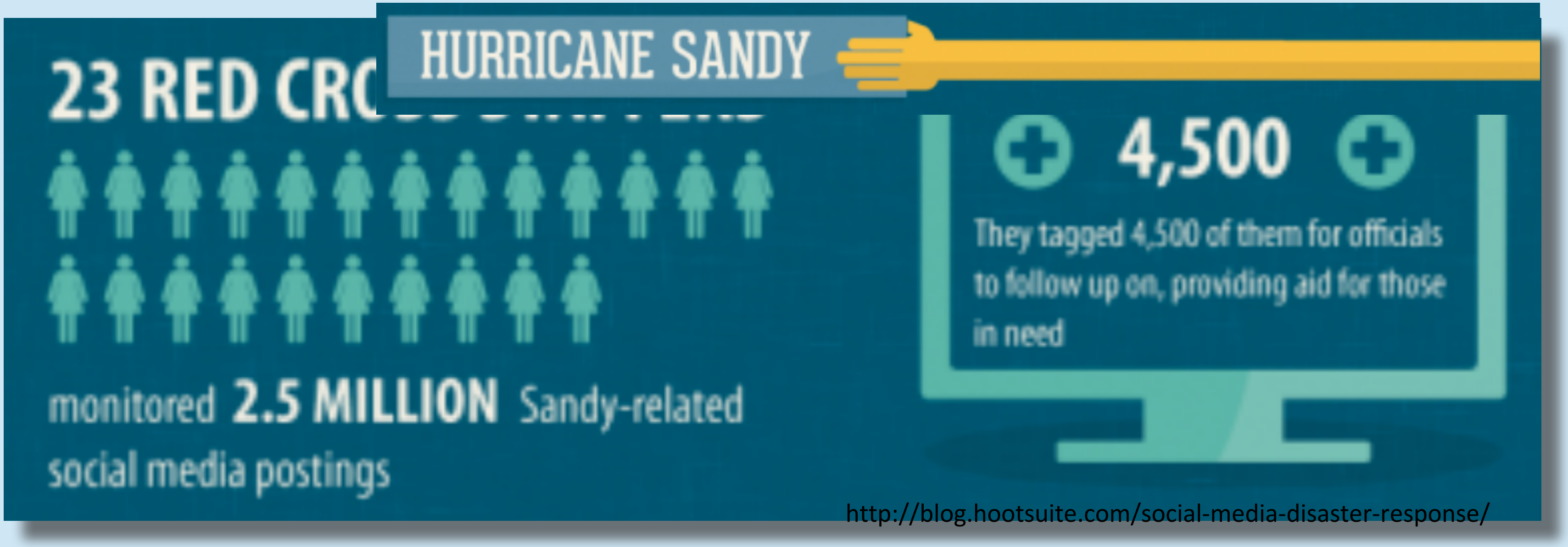
- Volume
- Velocity
- Variety

- High volume & velocity of messages:
  - Twitter has ~20 000 000 users per month
  - They write ~500 000 000 messages per day

- Massive variety:
  - Stock markets
  - Earthquakes
  - Social arrangements

Velocity

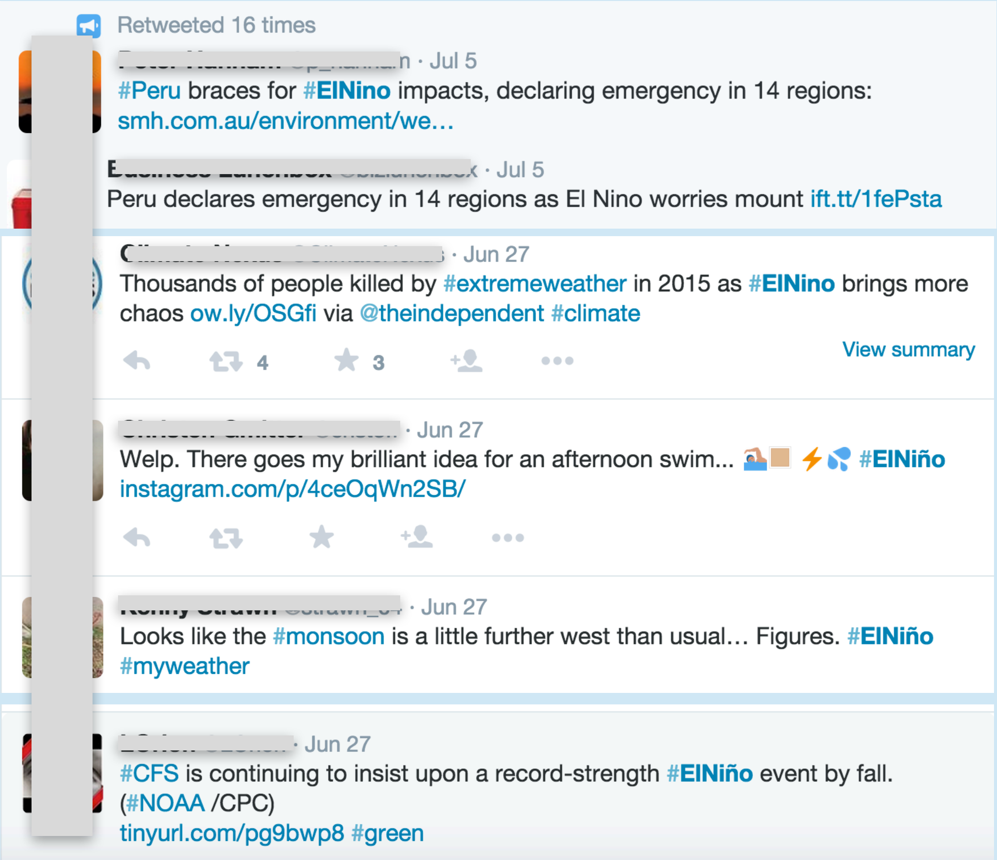
- During the Japan tsunami, over 1000 tweets were sent about the disaster every minute
- Processing such information streams manually is infeasible for disasters with heavy social media coverage and reporting


HURRICANE SANDY
23 RED CR
monitored 2.5 MILLION Sandy-related social media postings
4,500
They tagged 4,500 of them for officials to follow up on, providing aid for those in need

http://blog.hootsuite.com/social-media-disaster-response/

Informativeness

- Which messages are irrelevant or uninformative?
- The percentage of relevant and informative posts during crises varies a great deal
  - 10% during Missouri tornado in 2011
  - 65% during Australian bushfire in 2009



Retweeted 16 times

[redacted] · Jul 5
#Peru braces for #ElNino impacts, declaring emergency in 14 regions: smh.com.au/environment/we…

[redacted] · Jul 5
Peru declares emergency in 14 regions as El Nino worries mount ift.tt/1fePsta

[redacted] · Jun 27
Thousands of people killed by #extremeweather in 2015 as #ElNino brings more chaos ow.ly/OSGfi via @theindependent #climate

4 3 View summary

[redacted] · Jun 27
Welp. There goes my brilliant idea for an afternoon swim... 🏊🏽⛈⚡💦 #ElNiño instagram.com/p/4ceOqWn2SB/

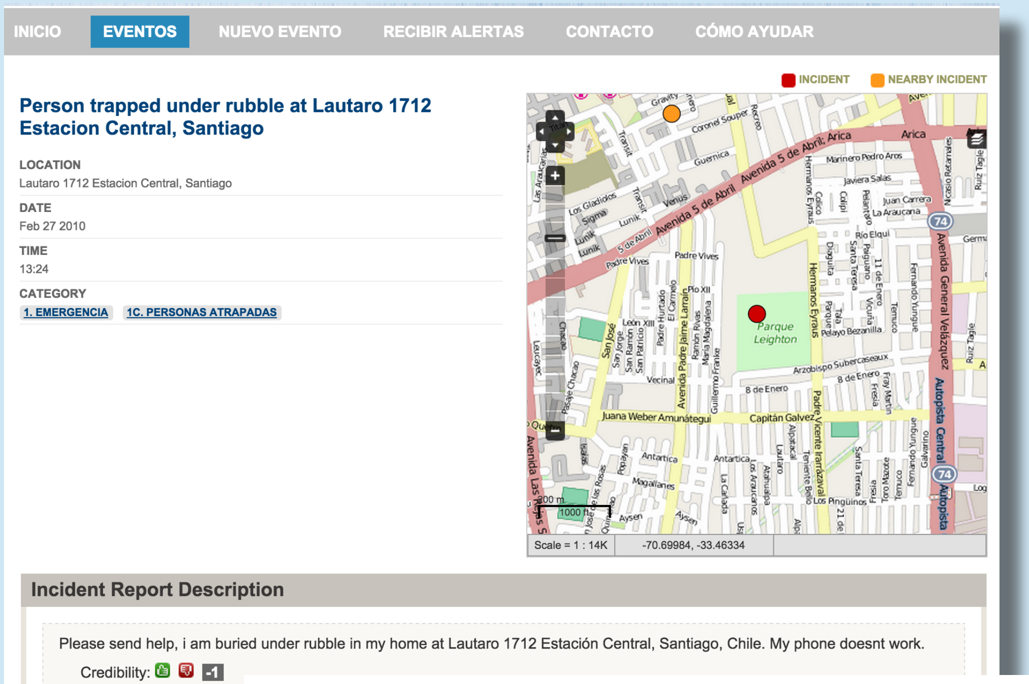
[redacted] · Jun 27
Looks like the #monsoon is a little further west than usual… Figures. #ElNiño #myweather

[redacted] · Jun 27
#CFS is continuing to insist upon a record-strength #ElNiño event by fall. (#NOAA /CPC)
tinyurl.com/pg9bwp8 #green

Content and Source Validity

- A study of the Ferguson civil unrest events in 2014 found that ~25% of the tweets were misinforming [Zubiaga, et al., 2015]

- Another study showed that ~10K tweets on Hurricane Sandy had fake pictures, and 0.3% of the users were behind 90% of the faked reports [Gupta, et al. 2013]



Person trapped under rubble at Lautaro 1712 Estacion Central, Santiago

LOCATION
Lautaro 1712 Estacion Central, Santiago

DATE
Feb 27 2010

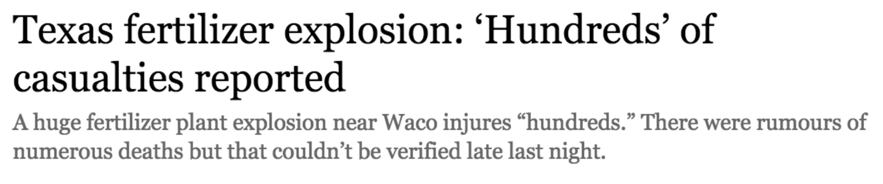
TIME
13:24

CATEGORY
1. EMERGENCIA 1C. PERSONAS ATRAPADAS

Incident Report Description

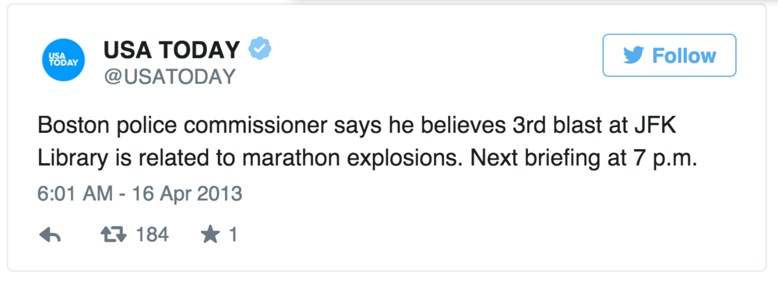
Please send help, i am buried under rubble in my home at Lautaro 1712 Estación Central, Santiago, Chile. My phone doesnt work.

Credibility: 👍 👎 -1



Texas fertilizer explosion: 'Hundreds' of casualties reported

A huge fertilizer plant explosion near Waco injures "hundreds." There were rumours of numerous deaths but that couldn't be verified late last night.



USA TODAY
@USATODAY

Boston police commissioner says he believes 3rd blast at JFK Library is related to marathon explosions. Next briefing at 7 p.m.

6:01 AM - 16 Apr 2013

184 ★ 1

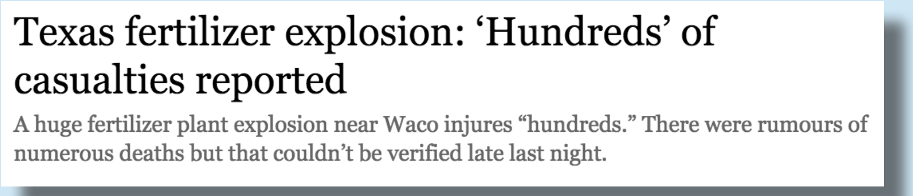
Content and Source Validity

- Many features have been studied to help identifying misinforming accounts
  - content features, information spread features, social network features, user profile, etc.

- Once found, what can we do about it?

After the 2013 Boston bombing, corrections to misinformation on twitter emerged but were **muted** in comparison to the propagation of misinformation [Starbird et al., 2014]

Rumours in Twitter during the Texas Fertilizer Plant Explosion on April 2013 did get corrected, but not fast and frequently enough to ensure that **accurate information** is shared within the community [Diver 2014]

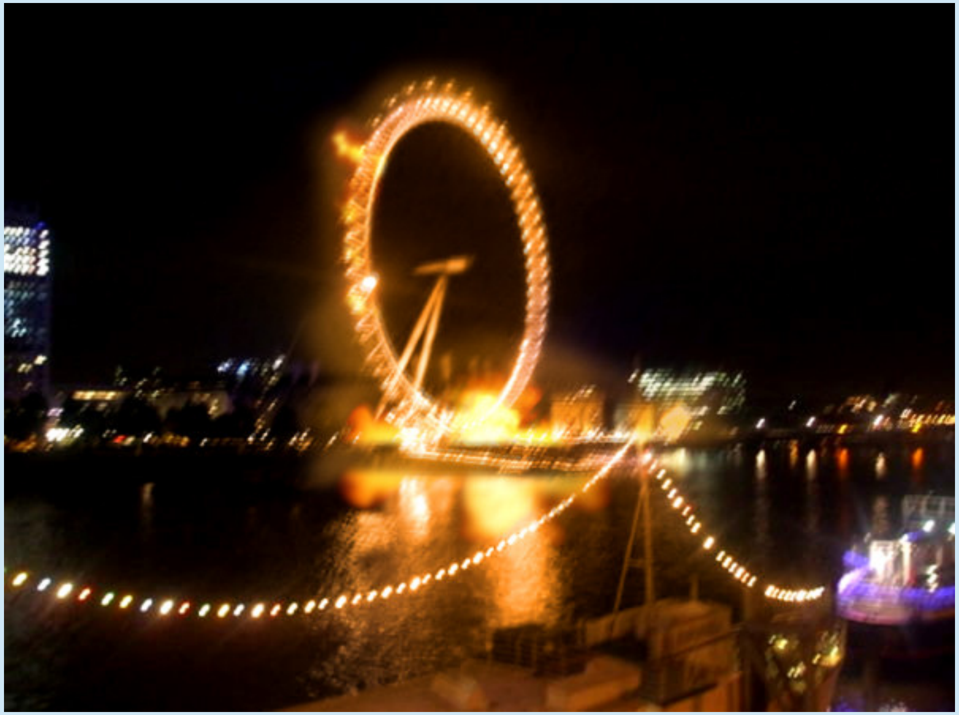


Texas fertilizer explosion: 'Hundreds' of casualties reported

A huge fertilizer plant explosion near Waco injures "hundreds." There were rumours of numerous deaths but that couldn't be verified late last night.

Oh my God!
This can't be happening at London Eye!

- *baconbkk*: This pic is not real. It is a photoshop giggle.
- *mactavish*: It's not. http://www.channel.com/news/london-riots-interactive-timeline-map



Problems of veracity in social media

- Most current rumour analysis has to be done manually
- Rumours are challenging: some could take days, weeks or even months to die out
- Ill-meaning humans can currently outsmart computers and appear completely genuine
- It's crucial for e.g. journalists, emergency services and people seeking medical information to know what's really true
- To combat this, we can draw on:
    - **NLP** to understand what's actually being said, resolve ambiguity etc.
    - **web science**: using a priori knowledge from Linked Data
    - **social science**: who spread the rumour, why and how
    - **information visualisation**: visual analytics

4 main kinds of rumour

- **uncertain information** or **speculation**
    - Greece will leave the Eurozone
- **disputed information** or **controversy**
    - aluminium causes Alzheimer’s
- **misinformation**
    - misrepresentation and quoting out of context
- **disinformation**
    - Obama is a Muslim

Using NLP to deal with veracity

- Tweets containing swearing and with poor grammar/spelling and little punctuation are likely to be real in a life-or-death scenario
- During an emergency, carefully worded tweets in journalistic style are less likely to be real tweets by eyewitnesses
- On the other hand, tweets containing valid medical information (as opposed to snake oil) are more likely to be written in good English

TamilNet reported that a second navy vessel had been sunk.

The Sri Lankan military denies that a second navy vessel had been sunk.

Other challenges of social media

- **Strongly temporal and dynamic**:
  - temporal information (e.g. post timestamp) can be combined with opinion mining, to examine the volatility of attitudes towards topics over time (e.g. gay marriage).
- **Exploiting social context**: who is the user connected to? How frequently do they interact?
  - Derive automatically semantic models of social networks, measure user authority, cluster similar users into groups, as well as model trust and strength of connection
- **Implicit information about the user**: research on recognising gender, location, and age of Twitter users.
  - Helpful for generating opinion summaries by user demographics

Social Media Sites

Twitter, LinkedIn, Facebook etc. have different properties

- Twitter has varied uptake per country:
    - Low in China (often censored, local competitor – Weibo)
    - Low in Denmark, Germany (Facebook is preferred)
    - Medium in UK, though often complementary to Facebook
    - High in USA
- Networks have common themes:
    - Individuals as nodes in a common graph
    - Relations between people
    - Sharing and privacy restrictions
    - No curation of content
    - Multimedia posting and re-posting

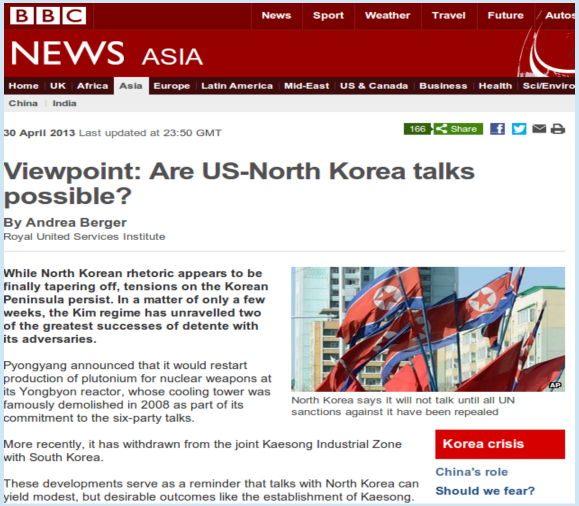
Linguistic challenges imposed by social media

- **Language:** social media typically exhibits very different language style
  - Solution: train specific language processing components
- **Relevance**: topics and comments can rapidly diverge.
  - Solution: train a classifier or use clustering techniques
- **Lack of context**: hard to disambiguate entities
  - Solution: data aggregation, metadata, entity linking techniques

Analysing language in social media is hard

- *Grundman:politics makes #climatechange scientific issue,people don’t like knowitall rational voice tellin em wat 2do*
- *@adambation Try reading this article , it looks like it would be really helpful and not obvious at all. http://t.co/mo3vODoX*
- *Want to solve the problem of #ClimateChange? Just #vote for a #politician! Poof! Problem gone! #sarcasm #TVP #99%*
- *Human Caused #ClimateChange is a Monumental Scam! http://www.youtube.com/watch?v=LiX792kNQeE … F**k yes!! Lying to us like MOFO's Tax The Air We Breath! F**k Them!*

NLP Pipelines



30 April 2013 Last updated at 23:50 GMT

Viewpoint: Are US-North Korea talks possible?

By Andrea Berger
Royal United Services Institute

While North Korean rhetoric appears to be finally tapering off, tensions on the Korean Peninsula persist. In a matter of only a few weeks, the Kim regime has unravelled two of the greatest successes of detente with its adversaries.

North Korea says it will not talk until all UN sanctions against it have been repealed

Pyongyang announced that it would restart production of plutonium for nuclear weapons at its Yongbyon reactor, whose cooling tower was famously demolished in 2008 as part of its commitment to the six-party talks.

More recently, it has withdrawn from the joint Kaesong Industrial Zone with South Korea.

These developments serve as a reminder that talks with North Korea can yield modest, but desirable outcomes like the establishment of Kaesong.

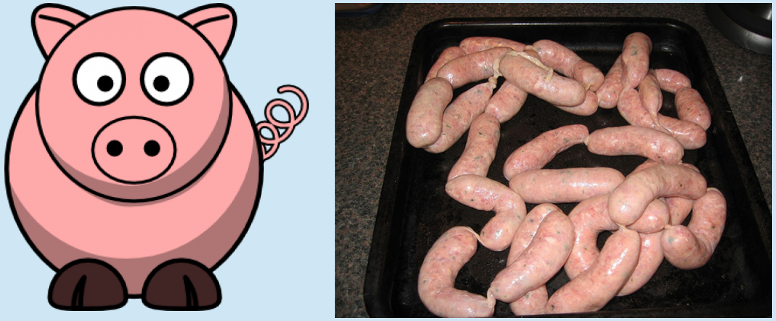
Korea crisis

China's role

Should we fear?

Text

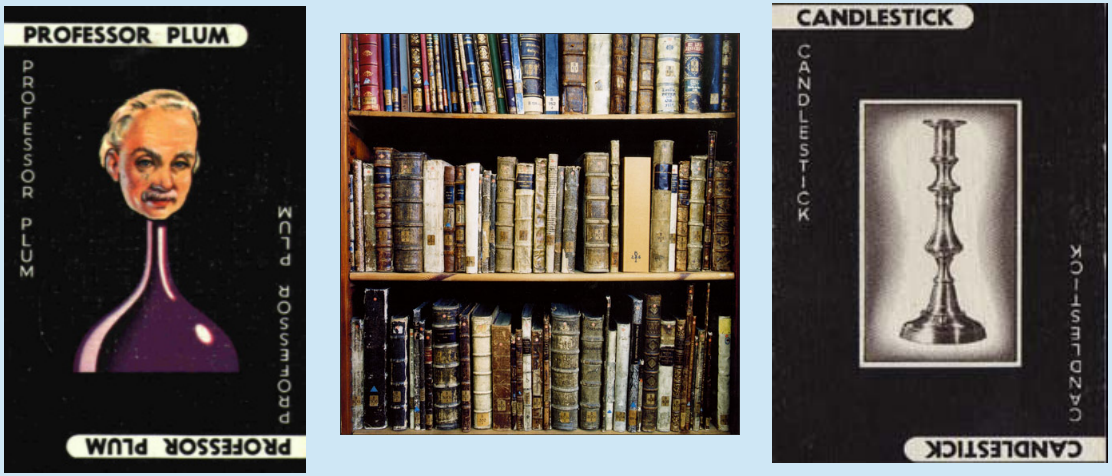
Language ID



Tokenisation

Part of speech tagging

Typical annotation pipeline



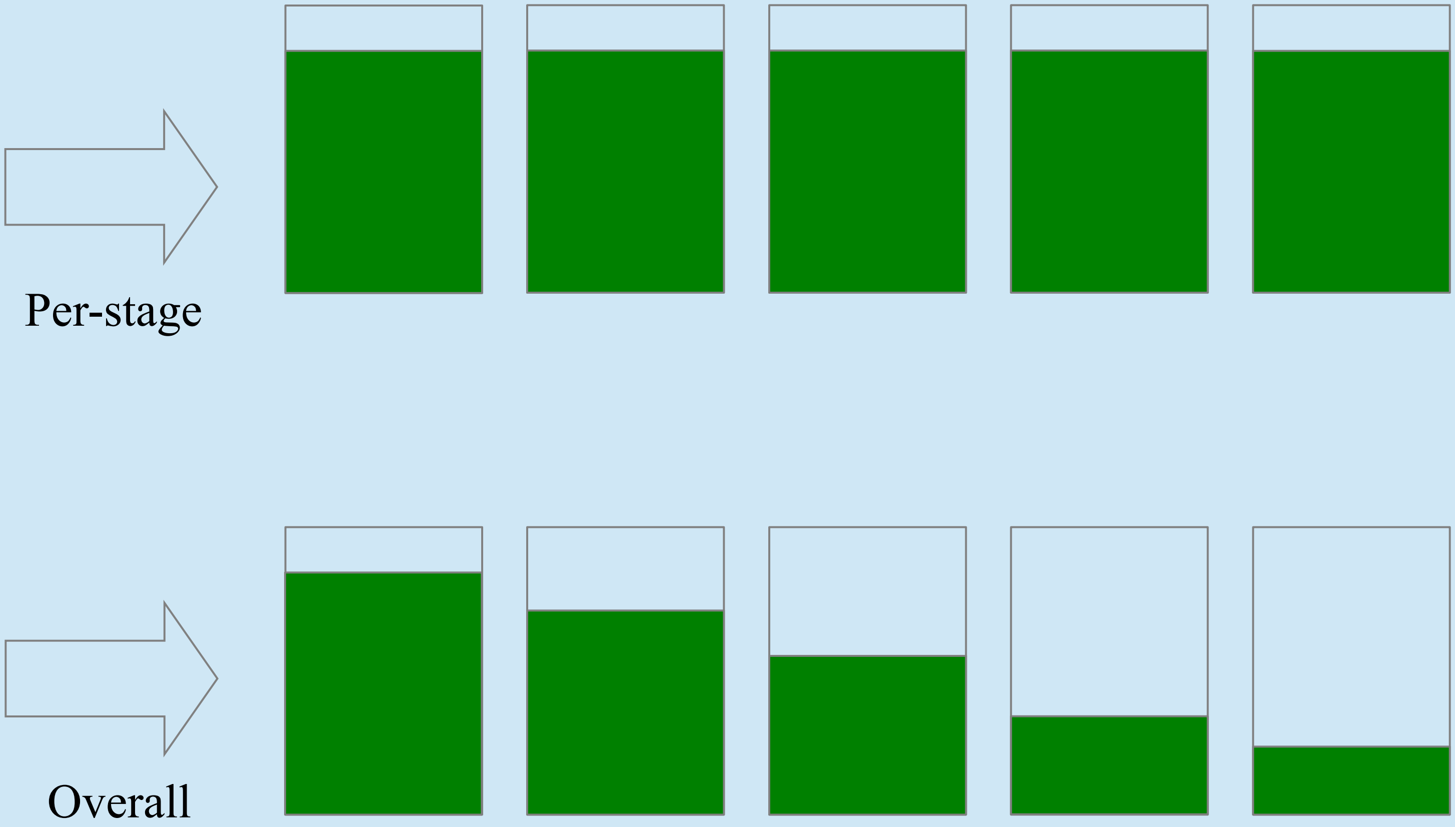
Named entity recognition

Linking entities

dbpedia.org/resource/.....
Michael_Jackson
Michael_Jackson_(writer)

Pipelines for tweets

- Errors have a cumulative effect



Good performance is important at each stage

Language ID: example

- Given a text, determine which language it is written in

Newswire:

The Jan. 21 show started with the unveiling of an impressive three-story castle from which Gaga emerges. The band members were in various portals, separated from each other for most of the show. For the next 2 hours and 15 minutes, Lady Gaga repeatedly stormed the moveable castle, turning it into her own gothic Barbie Dreamhouse .

Twitter:

LADY GAGA IS BETTER THE 5th TIME OH BABY(:

je bent Jacques cousteau niet die een nieuwe soort heeft ontdekt, het is duidelijk, ze bedekken hun gezicht. Get over it

I'm at 地铁望京站 Subway Wangjing (Beijing) http://t.co/KxHzYm00

RT @TomPIngram: VIVA LAS VEGAS 16 - NEWS #constantcontact http://t.co/VrFzZaa7

Language ID: issues

- Accuracy on microblogs: 89.5% vs on formal text: 99.4%

What general problems are there in identifying language in social media?

- Switching language mid-text;
- Non-lexical tokens (URLs, hashtags, usernames, retweets,...)
- Small "samples": document length has a big impact
- Dysfluencies and fragments reduce n-gram match likelihoods;
- Large number of potential languages: possibly no training data

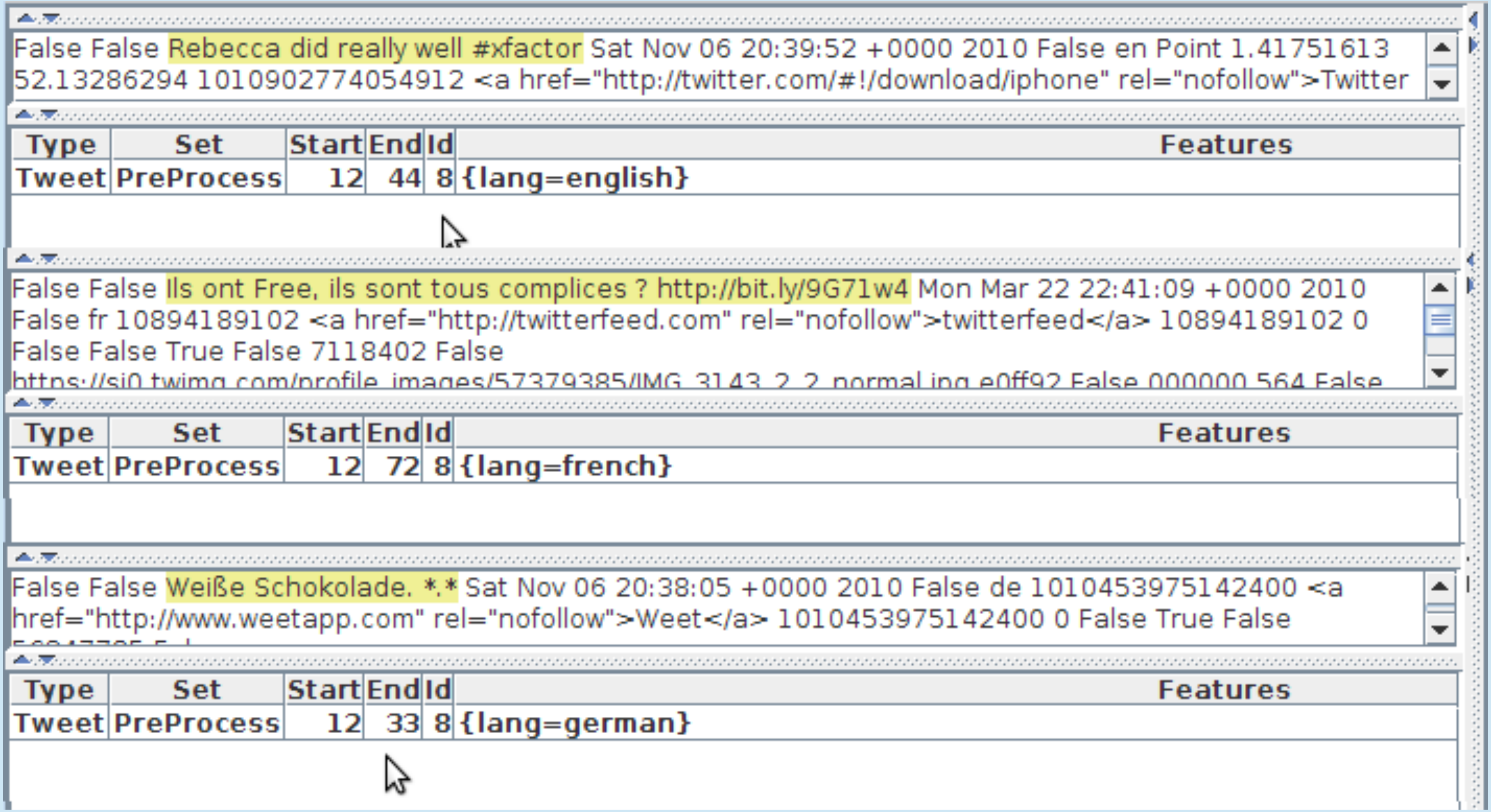
Social media introduces new sources of information.

- Metadata:
  spatial information (from profile, from GPS);
  language information (default English is left on far too often).
- Emoticons:
  :) vs. ^_^
  cu vs. 88

Language identification is tricky

- Language identification tools such as TextCat need a decent amount of text (around 20 words at least)
- But Twitter has an average of only 10 tokens/tweet
- Noisy nature of the words (abbreviations, misspellings).
- Due to the length of the text, we can make the assumption that one tweet is written in only one language (this isn't always the case though)
- We have adapted the TextCat language identification plugin
- Provided fingerprints for 5 languages: DE, EN, FR, ES, NL
- You can extend it to new languages easily

Language detection examples



False False Rebecca did really well #xfactor Sat Nov 06 20:39:52 +0000 2010 False en Point 1.41751613 52.13286294 1010902774054912 <a href="http://twitter.com/#!/download/iphone" rel="nofollow">Twitter

Type	Set	Start	End	Id	Features
Tweet	PreProcess	12	44	8	{lang=english}

False False Ils ont Free, ils sont tous complices ? http://bit.ly/9G71w4 Mon Mar 22 22:41:09 +0000 2010 False fr 10894189102 <a href="http://twitterfeed.com" rel="nofollow">twitterfeed</a> 10894189102 0 False False True False 7118402 False https://si0.twimg.com/profile_images/57379385/IMG_3143_2_2_normal.jpg e0ff92 False 000000 564 False

Type	Set	Start	End	Id	Features
Tweet	PreProcess	12	72	8	{lang=french}

False False Weiße Schokolade. *.* Sat Nov 06 20:38:05 +0000 2010 False de 1010453975142400 <a href="http://www.weetapp.com" rel="nofollow">Weet</a> 1010453975142400 0 False True False

Type	Set	Start	End	Id	Features
Tweet	PreProcess	12	33	8	{lang=german}

Tokenisation

- Plenty of “unusual”, but very important tokens in social media:
  - @Apple – mentions of company/brand/person names
  - #fail, #SteveJobs – hashtags expressing sentiment, person or company names
  - :-(, :-), :-P – emoticons (punctuation and optionally letters)
  - URLs
- Tokenisation is crucial for entity recognition and opinion mining
- Accuracy of standard tokenisers only 80% on tweets

Example

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

- Tokenising on white space doesn't work that well:
- Nike and Apple are company names, but if we have tokens such as #nike and @Apple, this will make the entity recognition harder, as it will need to look at sub-token level
- Tokenising on white space and punctuation characters doesn't work well either: URLs get separated (http, nikeplus), as are emoticons and email addresses

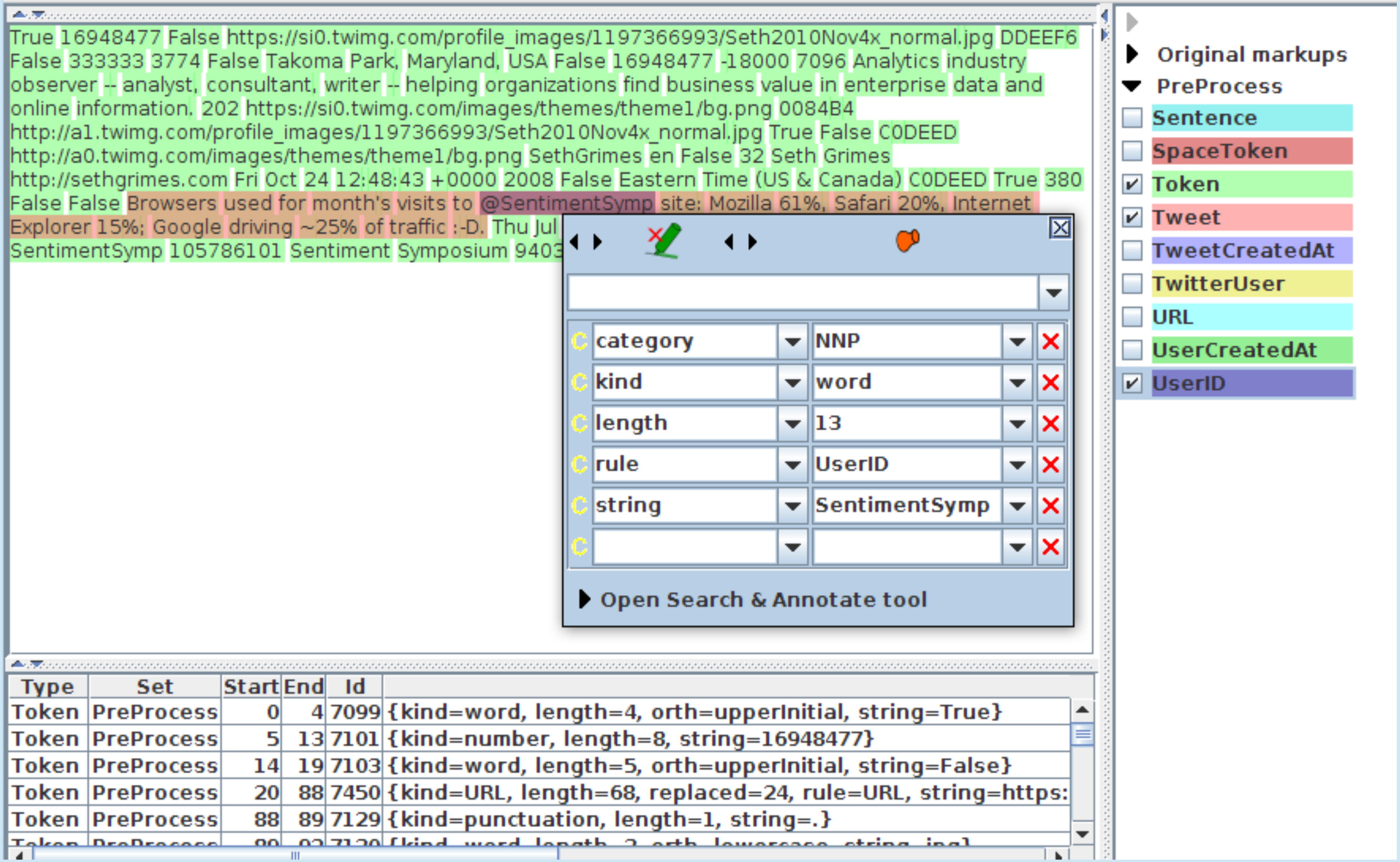
Tokenisation: issues

- Improper grammar, e.g. apostrophe usage:
  - doesn't → does n't
  - doesnt → doesn’t
  - Introduces previously-unseen tokens
- Smileys and emoticons
  - I <3 you → I & lt ; you
  - This piece ;,,(so emotional → This piece ; , , (so emotional
  - Loss of information (sentiment)
- Punctuation for emphasis
  - *HUGS YOU**KISSES YOU* → * HUGS YOU**KISSES YOU *
- Words run together / skip
  - I wonde rif Tsubasa is okay..

The TwitIE Tokeniser

- Treat RTs and URLs as 1 token each
- #nike is two tokens (# and nike) plus a separate annotation Hashtag covering both. Same for @mentions -> UserID
- Capitalisation is preserved, but an orthography feature is added: all caps, lowercase, mixCase
- Date and phone number normalisation, lowercasing, and emoticons are optionally done later in separate modules
- Consequently, tokenisation is faster and more generic
- Also, more tailored to our NER module

GATE Twitter Tokeniser: An Example



True 16948477 False https://si0.twimg.com/profile_images/1197366993/Seth2010Nov4x_normal.jpg DDEEF6 False 333333 3774 False Takoma Park, Maryland, USA False 16948477 -18000 7096 Analytics industry observer -- analyst, consultant, writer -- helping organizations find business value in enterprise data and online information. 202 https://si0.twimg.com/images/themes/theme1/bg.png 0084B4 http://a1.twimg.com/profile_images/1197366993/Seth2010Nov4x_normal.jpg True False C0DEED http://a0.twimg.com/images/themes/theme1/bg.png SethGrimes en False 32 Seth Grimes http://sethgrimes.com Fri Oct 24 12:48:43 +0000 2008 False Eastern Time (US & Canada) C0DEED True 380 False False Browsers used for month's visits to @SentimentSymp site: Mozilla 61%, Safari 20%, Internet Explorer 15%; Google driving ~25% of traffic :-D. Thu Jul SentimentSymp 105786101 Sentiment Symposium 9403

category	NNP
kind	word
length	13
rule	UserID
string	SentimentSymp

Type	Set	Start	End	Id	
Token	PreProcess	0	4	7099	{kind=word, length=4, orth=upperInitial, string=True}
Token	PreProcess	5	13	7101	{kind=number, length=8, string=16948477}
Token	PreProcess	14	19	7103	{kind=word, length=5, orth=upperInitial, string=False}
Token	PreProcess	20	88	7450	{kind=URL, length=68, replaced=24, rule=URL, string=https:
Token	PreProcess	88	89	7129	{kind=punctuation, length=1, string=.}
Token	PreProcess	89	92	7130	{kind=word, length=3, orth=lowercase, string=ing}

Analysing Hashtags

What's in a hashtag?

- Hashtags often contain smushed words
  - #SteveJobs
  - #CombineAFoodAndABand
  - #southamerica
- For NER we want the individual tokens so we can link them to the right entity

- For opinion mining, individual words in the hashtags often indicate sentiment, sarcasm etc.
  - #greatidea
  - #worstdayever

How to analyse hashtags?

- Camelcasing makes it relatively easy to separate the words, using an adapted tokeniser, but many people don't bother
- We use a simple approach based on dictionary matching the longest consecutive strings, working L to R
  - #lifeisgreat -> #-life-is-great
  - #lovinglife -> #-loving-life
- It's not foolproof, however
  - #greatstart -> #-greats-tart
- To improve it, we could use contextual information, or we could restrict matches to certain POS combinations (ADJ+N is more likely than ADJ+V)

Tweet Normalisation

- "RT @Bthompson WRITEZ: @libbyabrego honored?! Everybody knows the libster is nice with it...lol...(thankkkks a bunch;))"
- OMG! I'm so guilty!!! Sprained biibii's leg! ARGHHHHHH!!!!!!
- Similar to SMS normalisation
- For some components to work well (POS tagger, parser), it is necessary to produce a normalised version of each token
- BUT uppercasing, and letter and exclamation mark repetition often convey strong sentiment
- Therefore some choose not to normalise, while others keep both versions of the tokens

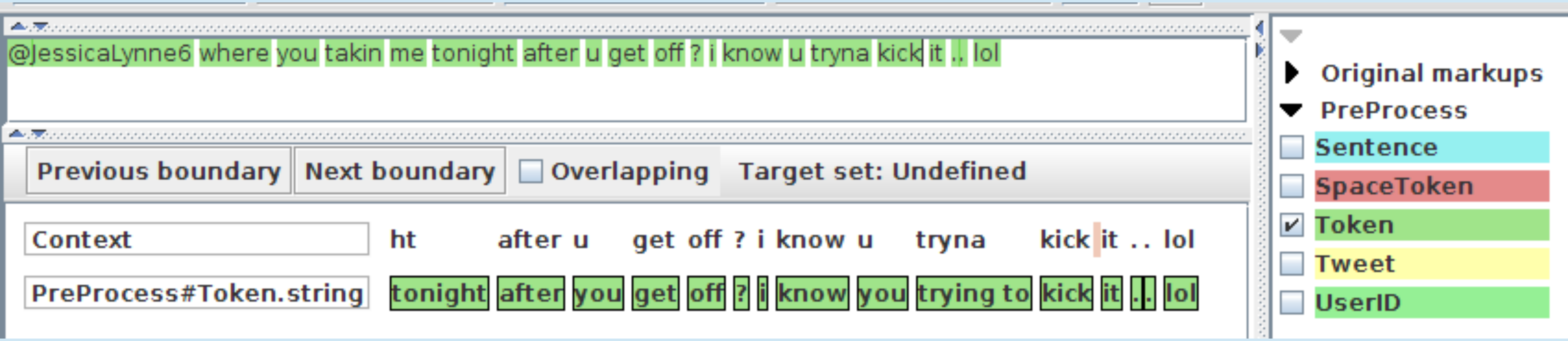
Lexical normalisation

- Two classes of word not in dictionary
  - 1. Mis-spelled dictionary words
  - 2. Correctly-spelled, unseen words (e.g. foreign surnames)
  - Problem: Mis-spelled unseen words
- First challenge: separate out-of-vocabulary and in-vocabulary
- Second challenge: fix mis-spelled IV words
  - Edit distance (e.g. Levenshtein): count character adds, removes

  zseged → szeged (distance = 2)
  - Pronunciation distance (e.g. double metaphone):

  YEEAAAHHH → yeah
- Need to set bounds on these, to avoid over-normalising OOV words

A normalised example



@JessicaLynne6 where you takin me tonight after u get off ? i know u tryna kick it .. lol

Context	ht after u get off ? i know u tryna kick it .. lol
PreProcess#Token.string	tonight after you get off ? i know you trying to kick it .. lol

- Original markups
- PreProcess
  - □ Sentence
  - □ SpaceToken
  - ☑ Token
  - □ Tweet
  - □ UserID

- Normaliser currently based on spelling correction and some lists of common abbreviations
- Outstanding issues:
  - Some abbreviations which span token boundaries (e.g. gr8, do n’t) difficult to handle
  - Capitalisation and punctuation normalisation

Part-of-speech tagging: example

Many unknowns:

- Music bands: **Soulja Boy | TheDeAndreWay.com in stores Nov 2, 2010**
- Places: **#LB #news: Silverado Park Pool Swim Lessons**

Capitalisation way off

- **@thewantedmusic on my tv :) aka derek**
- **last day of sorting pope visit to birmingham stuff out**
- **Don't Have Time To Stop In??? Then, Check Out Our Quick Full Service Drive Thru Window :)**

Part-of-speech tagging: example

Slang

- **~HAPPY B-DAY TAYLOR !!! LUVZ YA~**

Orthographic errors

- **dont even have homwork today, suprising?**

Dialect

- **Ey yo wen u gon let me tap dat**
  - ***Can I have a go on your iPad?***

Part-of-speech tagging: issues

Low performance

- Using in-domain training data, per token:
  SVMTool 77.8%, TnT 79.2%, Stanford 83.1%
- Whole-sentence performance: best was 10%
- Best performance on newswire about 55-60%

Problems on unknown words – this is a good target set to get better performance on

- 1 in 5 words completely unseen
- 27% token accuracy on this group

Tackling the problems: unseen words in tweets

- Majority of non-standard orthographies can be corrected with a gazetteer:

Vids	→	videos
cussin	→	cursing
hella	→	very

- No need to bother with e.g. Brown clustering
- 361 entries give 2.3% token error reduction

Part-of-speech tagging: solutions

- Not much training data is available, and it is expensive to create
- Plenty of unlabelled data available – enables e.g. bootstrapping
- Existing taggers algorithmically different, and use different tagsets with differening specificity

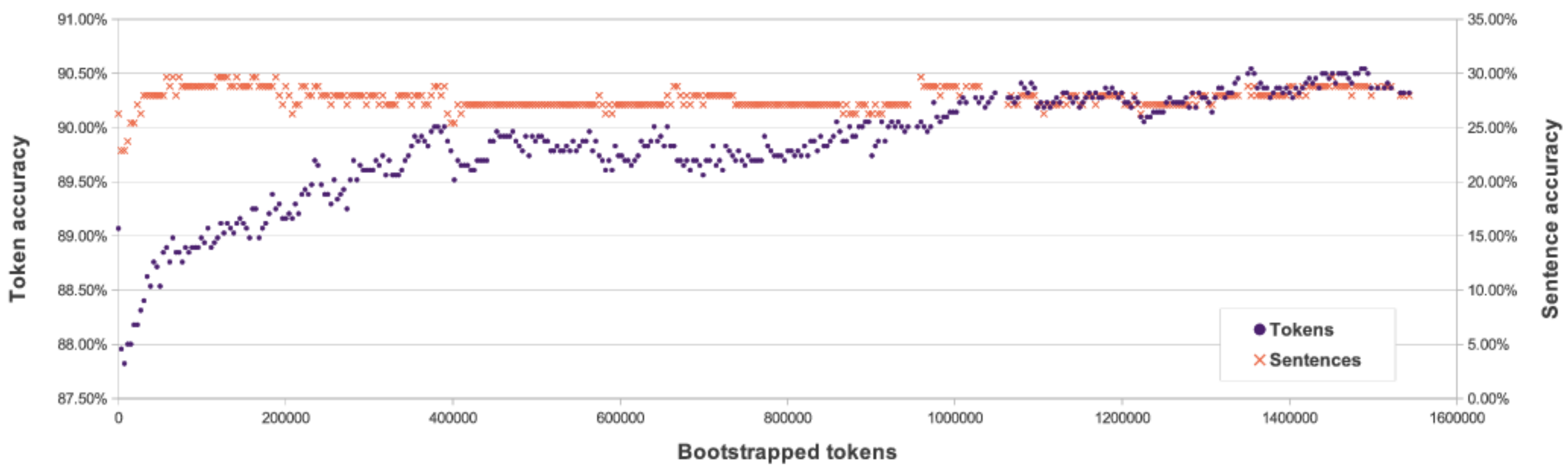
    - CMU tag **R** (adverb) → PTB (**WRB**,**RB**,**RBR**,**RBS**)
    - CMU tag **!** (interjection) → PTB (**UH**)

Part-of-speech tagging: solutions

- Label unlabelled data with taggers and accept tweets where tagger votes never conflict

- Lebron_^ + Lebron_NNP → OK, Lebron_NNP
- books_N + books_VBZ → Fail, reject whole tweet

Token accuracy: 88.7% sentence accuracy: 20.3%



Part-of-speech tagging: other solutions

- Non-standard spelling, through error or intent, is often observed in twitter – but not newswire

- Model words using Brown clustering and word representations (Turian 2010)
- Input dataset of 52M tweets as distributional data
- Use clustering at 4, 8 and 12 bits; effective at capturing lexical variations
  - E.g. cluster for "tomorrow": 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

- Data and features used to train CRF. Reaches 41% token error reduction over Stanford tagger.

Lack of context causes ambiguity

Branching out from Lincoln park after dark ... Hello Russian Navy, it's like the same thing but with glitter!

??

Getting the NEs right is crucial

Branching out from Lincoln park after dark ... Hello Russian Navy, it's like the same thing but with glitter!

Strategies for NER on social media

- Parsing is pretty terrible still on tweets, so not recommended
- Ambiguity is much more frequent, due to capitalisation etc., especially on first names
- “Individuals play the game, but teams beat the odds.”

Beating the Odds

- Depends a lot on your text and domain: what are the odds of first names appearing as common nouns vs proper nouns?
- will, may, autumn etc. are unlikely to be names if they occur on their own
- “bill” in a corpus of tweets about energy bills. Or ducks.
- We can't rely on POS tagging to be accurate
- Use more contextual information
- Semantic annotation can be useful to gain extra information about NEs, e.g. “will smith”

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
  - Match e.g. London - Londom
- Extended Gazetteer: has a number of parameters for matching prefixes, suffixes, initial capitalisation and so on
  - Match e.g. London - Londonnnn

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
  - 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 co-reference)

Resolving entity ambiguity

- We can resolve entity reference ambiguities with disambiguation techniques / linking to URI

  A plane just crashed in Paris. Two hundred French dead.
  - Paris (France), Paris (Hilton),or Paris (Texas)?
  - Match NEs in the text and assign a URI from all DBpedia matching instances
  - For ambiguous instances, calculate the disambiguation score (weighted sum of 3 metrics)
  - Select the URI with the highest score
- Try the demo at https://cloud.gate.ac.uk/shopfront/displayItem/yodie-en

Hands-on with TwitIE

- First we need to load the TWITIE plugin (green jigsaw icon)
- Scroll down to the Twitter plugin and select “Load now”.
- Click “Apply All” and then “Close”.
- Now right-click on “Applications”, select “Ready-made applications” and “TwitIE”
- Create another new corpus, name it “Tweets”
- Right-click on the corpus and select “populate from Twitter JSON”, selecting the file hands-on-materials/corpora/energy-tweets.json
- Run TwitIE on the corpus
- Look at the different annotations in the default annotation set
- To see Tokens in hashtags, use the Annotation Stack view

Extra hands-on:
What happens if you run ANNIE on tweets?

- Try running ANNIE on the tweets corpus and see how it differs from TwitIE
- Adventurous GATE users can change the name of the annotation set for one of the applications, and then run the Corpus QA or AnnotationDiff tool to compare the two sets of results more easily
- Adventurous GATE users can try comparing the applications on other corpora

Summary

- Very quick tour of some of the problems of text analysis on social media
- Some solutions proposed
- See TwitIE in action in GATE
- Next, we’ll look at sentiment analysis on social media