Linguistic preprocessing of social media

Leon Derczynski

General challenges

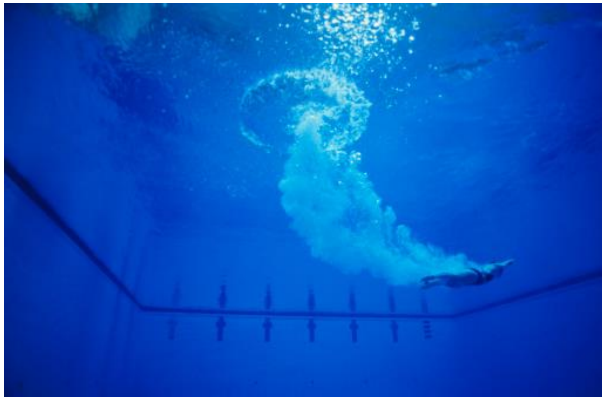
Common complaints we have about social media text:

- Documents are short;
- There are spelling errors;
- Words are ambiguous;
- Nonstandard / new lexical items;
- Nonstandard syntactic patterns.

The impact (or the cause?) of these complaints: Low performance of existing systems.

Maybe we need to re-train?

- Shortage of training data;
- Low-performance of existing techniques.



How can we characterise social media text?

What new techniques can help us process it?

Let's start at the deep end: Twitter text.*

* also – it's public and plentiful

Qualitative genre description

Great diversity in social media users, but they're not illiterate

- People want to represent their own dialects and accents (Jones 2010)
- They pick and choose from the entire stylistic repertoire of language (Tagliamonte 2008)
- Same literacy scores in standard and non-standard vocabulary users (Drouin 2009)

Emoticons have more than just an expressive function

- Pragmatic function, e.g. demonstrating a less stressed stance (Dresner 2010)
- Not just pictograms: phrasal abbreviations are also included – *smh*, *lol*
- Lexical items are made nonstandard through lengthening – *cooolll* (Brody 2011)

Social variables associated with certain transformations

- Slang is less inhibited in informal settings (Labov 1972)
- G-dropping mapped from speech to writing (Eisenstein 2010)
- Lexemes can have a spatial association within a language (Eisenstein 2011)

This socio-linguistic variation in social media highlights bias in existing resources

- Most corpora text was curated predominantly by working-age white men (Eisenstein 2013)
- Social media is not curated, so has different biases
- We have little data that is free from this demographic bias

From Eisenstein (2013)

Quantitative genre description

General style

- Twitter is more conservative and formal, less conversational than SMS and online chat;
- It still has a similar brevity to these mediums, but word choice is careful, with high density of lexical words (Halliday 2004);
- Tweets are used for sharing news or broadcasting personal status

Individual style:

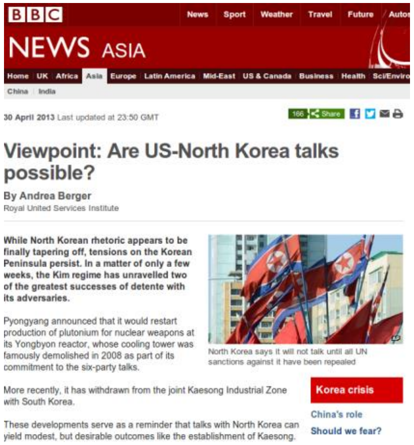
- Do users prefer individualistic style, or address a large audience? (Yates 1996)
- Users develop linguistically unique styles compared to other mediums;
- For example, both 1st and 3rd person pronouns are common, where other genres tend to stick to just one.
- Intensifier use indicates a younger audience - “really” vs. “very” (Ito 2003).

Temporal reference

- Are authors concerned with a certain timeframe? (past, present, future relative to timestamp)
- Temporal references are similar to SMS and online chat: no particular focus

From Hu et al. (2013)

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

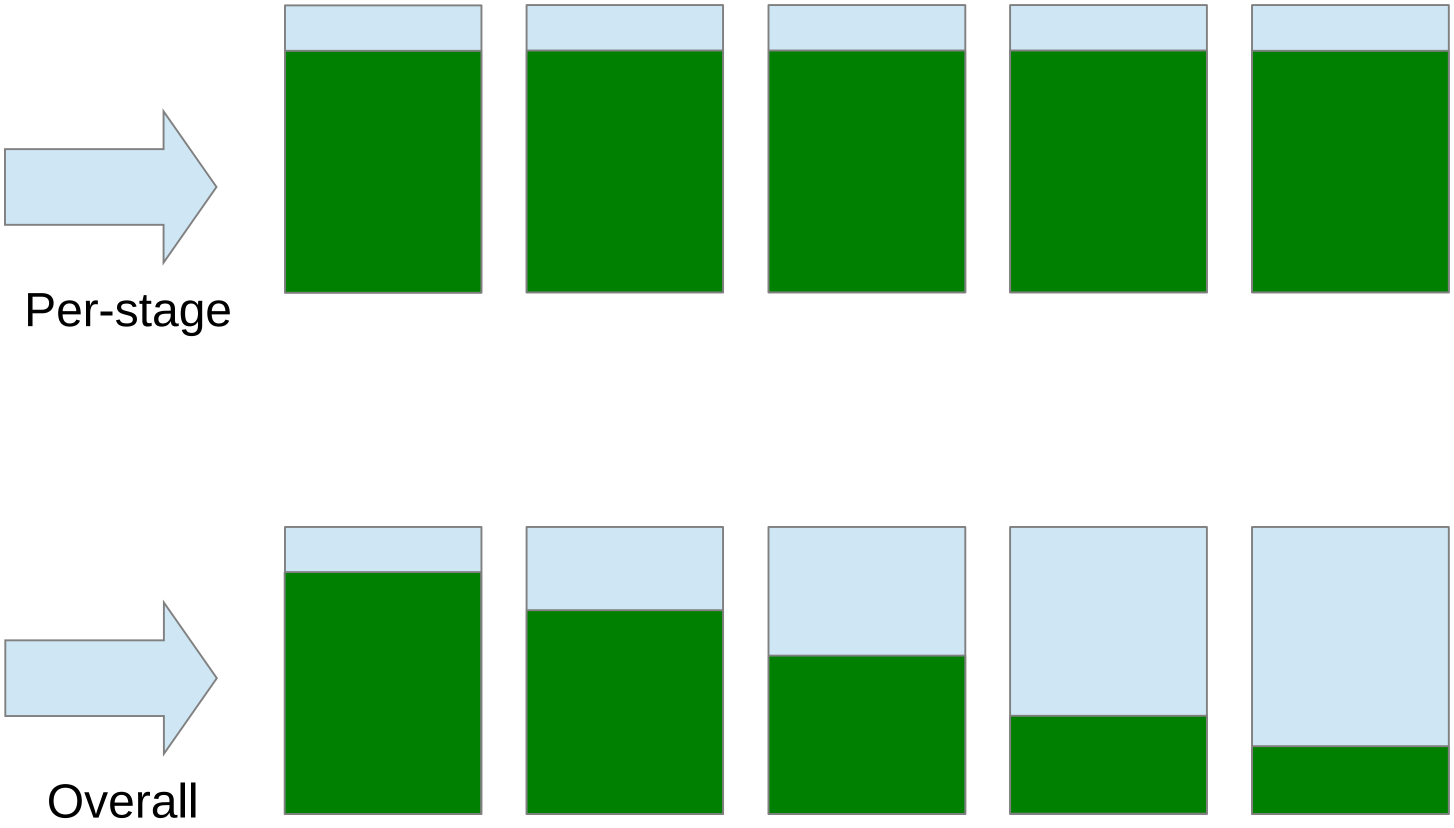
Language ID

Tokenisation

Part of speech tagging

Pipelines for tweets

- Errors have a cumulative effect



Good performance is important at each stage

Language ID: example

Task: given a text, determine which language it is intended to be.

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 .

Language ID: example

Task: given a text, determine which language it is intended to be.

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

General accuracy on microblogs:	89.5% (Preotiuc-Pietro 2012)
Compared to accuracy on formal text:	99.4% (Carter 2013)

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

- Switching language mid-text;
- Non-lexical tokens (URLs, hashtags, usernames, retweet/modified tweet indicators);
- Small “samples”: documents are fixed at 140 characters, and document length has a big impact on language identification;
- Dysfluencies and fragments reduce n-gram match likelihoods;
- Large (unknown) number of potential languages, some for which there will be no training data (Baldwin 2010).

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 ID: solutions

Carter et al. (2013) introduce semi-supervised priors to overcome short message problems:

- Author prior, using content of previous messages from the same author;
- Link prior, using text from any hyperlinks in the message;
- Mention prior, based on the author priors of other users mentioned in the message;
- Tag prior, gathering text in other messages sharing hashtags with the message;
- Conversation prior, taking content from messages in a conversation thread.

These priors individually help performance

- Author prior offers 50% error reduction, and is most helpful in five languages surveyed.
- Why? This prior will generate the most content – the others are conditional.

Combining priors leads to improved performance

- Different strategies help for different languages;
- Tried: voting, beam search, linear interpolation, beam confidence, lead confidence.
- Beam confidence (reducing prior weight when many languages close to most likely).

Tricky cases remain difficult, especially when languages mix

- Fluent multilingual posts; foreign named entities; misleading priors; language ambiguous

Language ID: solutions

Carter technique can be demanding

- Data may not be available: API limits, graph changes, deleted items, changed web pages
- Processing time: retrieving required information is slow
- Privacy concerns: somewhat invasive

Lui and Baldwin (2012) use information gain-based feature selection for transductive language ID

- Goal is to develop cross-domain language identification
- In-domain language identification is significantly easier than cross-domain
- Social media text is more like a mixture of small/personal domains than its own domain

The variety of data and sparsity of features makes selection important

- LD focuses on task-relevant features using information gain
- Features with a high LD score are informative about language, without being informative about domain
- Candidate features pruned before applying LD based on term frequency

Without training, the langid.py tool does better than other language ID systems on social media

- Consistent improvement over plain TextCat, LangDetect and CLD
- Limited to no training data available for the 97 target languages

Tokenisation: example

General accuracy on microblogs: 80%

Goal is to convert byte stream to readily-digestible word chunks

Word bound discovery is a *critical* language processing task

Newswire:

The LIBYAN AID Team successfully shipped these broadcasting equipment to Misrata last August 2011, to establish an FM Radio station ranging 600km, broadcasting to the west side of Libya to help overthrow Gaddafi's regime.

Twitter:

RT @JosetteSheeran: @WFP #Libya breakthru! We move urgently needed #food (wheat, flour) by truck convoy into western Libya for 1st time :D

@ojmason @encoffeedrinker But it was #nowthatcherisdead that was confusing (and not just to non-UK people!)

RT @Huddy85 : @Mz_Twilightxxx *kisses your asssneezes after* Lol**

Ima get you will.i.am NOTHING IS GONNA STAND IN MY WAY =)

Tokenisation: issues

Social media text is generally not curated, and typographical errors are common

Improper grammar, e.g. apostrophe usage:

- doesn't → does n't
- doesnt → doesnt
- 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..

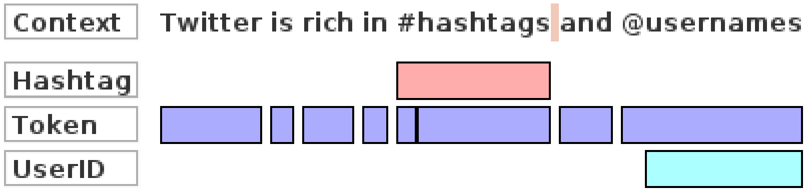
Tokenisation: solutions

O'Connor et al. (2010) apply a regular expression tokeniser to tweets, with the following adaptations:

- Treat #hashtags, @mentions, abbreviations, strings of punctuation, emoticons and UTF glyphs as single tokens
- Made available as “twokenizer” tool

Bontcheva et al. (2013) extend the Penn Treebank tool with twitter adaptations

- Layer multiple annotations on top of each other: Hashtags, Usernames



Twitter is rich in #hashtags and @usernames

- Normalisation maps frequent nonstandard spellings to standard
  - Via lookup dictionary (e.g. Han 2011); e.g. gonna → going to
  - Regular expressions for known smileys/emoticons to avoid splitting them

- Segmenting individual hashtags is possible (Maynard 2014)
  - #openaccess → # open access
  - #swankkkkk → # swan kkk k k ?

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 :)**

Slang

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

Orthographic errors

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

Dialect

- **Shall we go out for dinner this evening?**
- **Ey yo wen u gon let me tap dat**

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%; cf. SotA 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

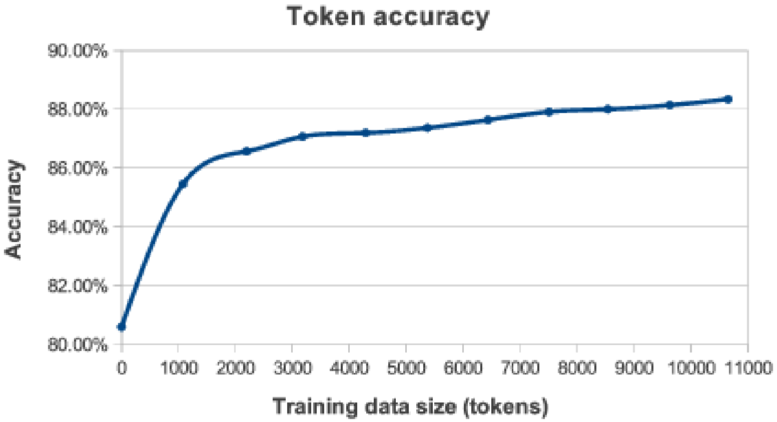
Errors on unknown words

- Gold standard errors (dank_UH je_UH → _FW) (Plank 2014)
- Training lacks IV words (Internet, bake)
- Pre-taggables (URLs, mentions, retweets)
- NN vs. NNP (derek_NN, Bed_NNP)
- Slang (LUVZ, HELLA, 2night)
- Genre-specific (unfollowing)
- Leftover tokenisation errors (ass**sneezes)
- Orthographic (suprising)

Part-of-speech tagging: issues

Insufficient data

- Ritter: 15K tokens, PTB, one annotator
- Foster: 14K tokens, PTB, low-noise
- CMU: 39K tokens, custom, narrow tagset



Unknown words fall roughly into two categories

- Standard token, non-standard orthography;
  - freinds
  - KHAAAANNNNNNN!
- Non-standard token, standard orthography
  - omg + bieber → omb
  - Huntingdon / Huntington

Part-of-speech tagging: solutions

Ritter et al. (2011) adapt to twitter by looking beyond newswire and modelling lexical variation

Extra resources include adapting standards to the genre and finding more & better data

- Extension of PTB tagset, with HT, USR, RT, and URL
- Inclusion of an IRC dataset (online chat; assumed similar to twitter; source of hashtag)
- Creation of a new Twitter corpus – 15K tokens, single annotator

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.

Part-of-speech tagging: solutions

Derczynski et al. (2013) extend the Ritter work, identifying three techniques for better performance

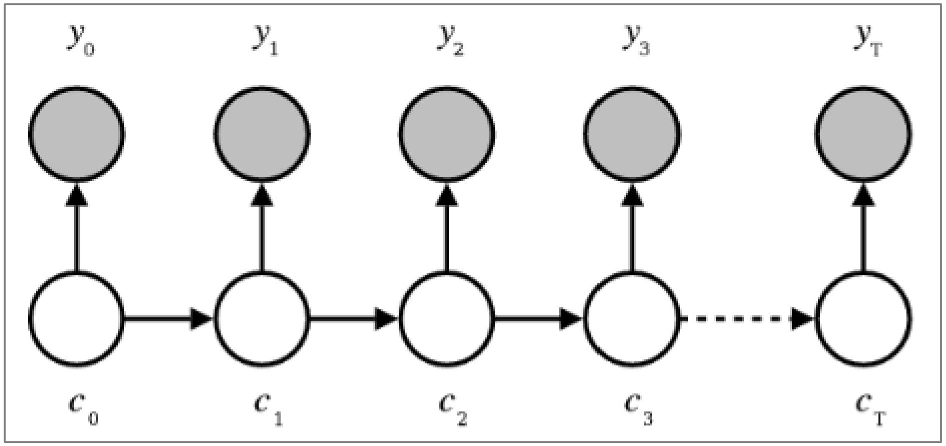
Unusual words still cause problems. How can these be better covered?

- Majority can be corrected via gazetteer: vids → videos, cussin → cursing, hella → very
- 361 entries give 2.3% token error reduction
- The rest can handled reasonably with word shape and contextual features
- Features include:
  - word prefix and suffix shapes
  - distribution of shape in corpus
  - shapes of neighbouring words

Part-of-speech tagging: solutions

For some tokens, we know the label with complete certainty

- Links, hashtags, user mentions, some smileys
- Some monosemous words, e.g. organisations, in gazetteers

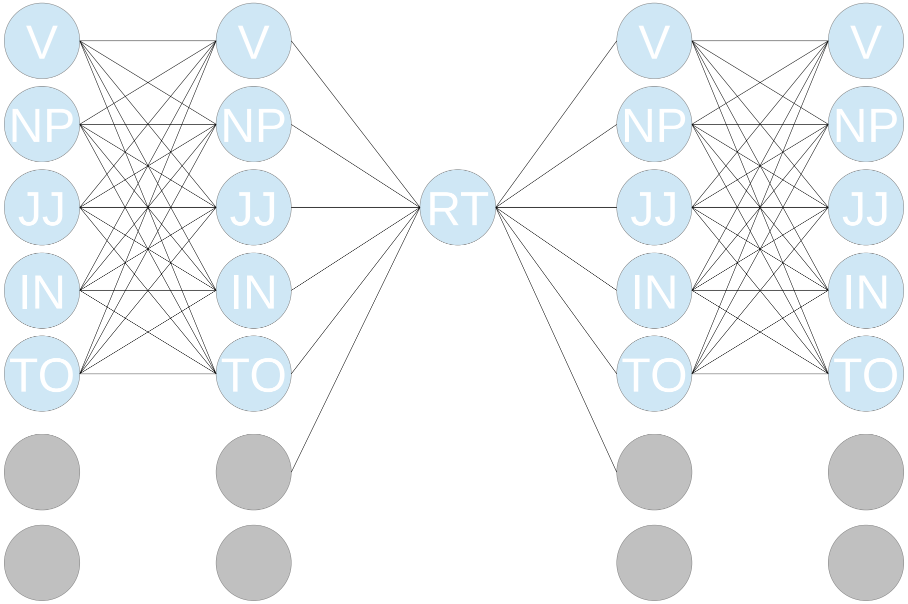


- We can fix the labels for these tokens.

Part-of-speech tagging: solutions

For some tokens, we know the label with complete certainty

- This allows us to prune the transition graph of labels in the sequence



- Because the graph is read in both directions, fixing the value of any label impacts whole tweet
- Setting label priors reduces token error 5.03%

Part-of-speech tagging: solutions

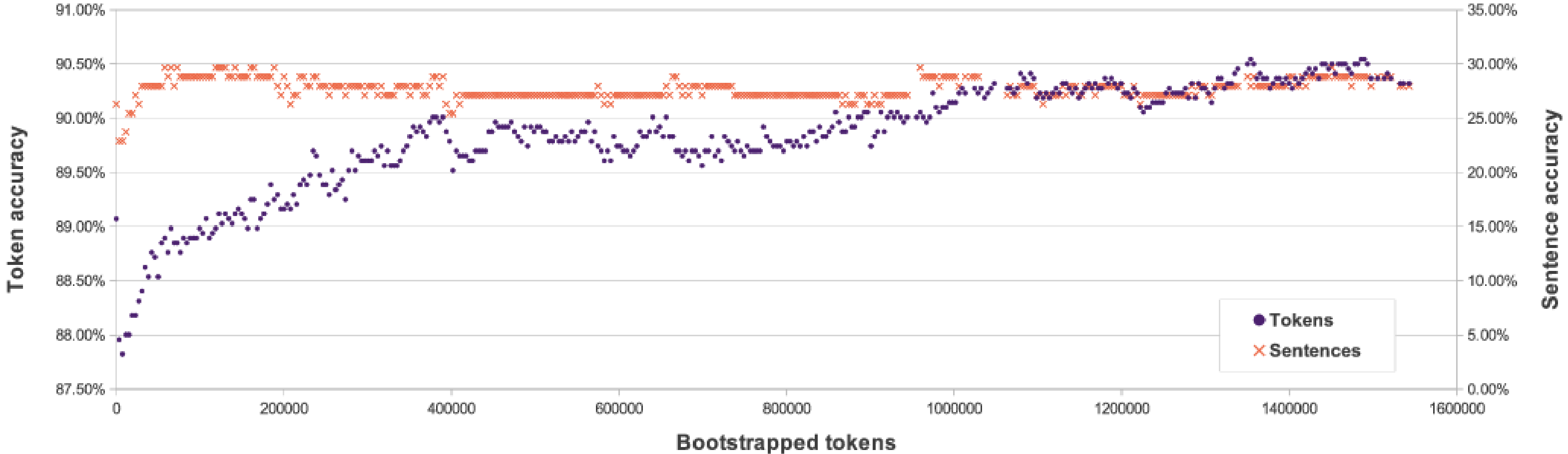
Not much training data is available, and it is expensive to create

- 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**)

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: solutions

Gimpel et al. (2011) adopt a holistic approach to PoS tagging

A tagset is created that adapts to the tokenisation issues already seen

- No splitting contractions; instead, combined forms added. {nn, nnp} x {vb, pos}
- New tags for twitter phenomena (#, @, ~ for RT, U for URL) and emoticons (E)
- Choose to annotate mid-sentence hashtags as other parts of speech
- Leads to new corpus, tokenised and tagged: 39K tokens, 0.92 IAA

Twitter-specific features used with CRF

- Orthographic, detecting fixed-format tokens
- Frequently-capitalised tokens are collected, to overcome capitalisation inconsistency
- Prior tag distribution taken from PTB (including Brown) as soft prior
- Distributional similarity taken from 1.9M unlabelled tweets, looking one ahead & behind
- Phonetic representations are taken using metaphone, and compared with tag distributions in PTB of words sharing the metaphone key

Owoputi et al. (2013) extend using word clusters, proper name gazetteers, and regularisation

Final accuracy: 93.2% token-level, ~22% sentence-level

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 domain / genre, and create customised tools and techniques
- Some success in language ID (Carter 2013), PoS tagging (Gimpel 2011), NER (Ritter 2011)

User adaptation

- A “third way”: social media as a whole is not a distinct genre or in need of repair
- Suggested by Eisenstein 2013, Baldwin 2010, Hu 2013
- Instead, composed of many users each with their own styles (cf. AP guidelines)
- Incorporating per-user models offers insights into communications there (cf. Carter)

Bibliography

Baldwin, Lui 2010. Language Identification: The Long and the Short of the Matter. NAACL

Bontcheva, Derczynski, Funk, Greenwood, Maynard, Aswani 2013. TwitIE: An open-source information extraction pipeline for microblog text. RANLP

Brody, Diakopoulos 2011. Cooooooooooooollllllllllllll!!!!!!!!!!!!: using word lengthening to detect sentiment in microblogs. EMNLP

Carter, Weerkamp, Sagkias 2013. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. JLRE

Derczynski, Maynard, Aswani, Bontcheva 2013a. Microblog-genre noise and impact on semantic annotation. Hypertext

Derczynski, Ritter, Clark, Bontcheva 2013b. Twitter part-of-speech tagging for all: overcoming sparse and noisy data. RANLP

Dresner, Herring 2010. Functions of the non-verbal in CMC: Emoticons and illocutionary force. Communication Theory, 20(3):249

Drouin, Davis 2009. R u txting? Is the use of text speak hurting your literacy? Journal of Literacy Research, 41(1):46

Eisenstein, O'Connor, Smith and Xing 2010. Discovering sociolinguistic associations with structured sparsity. ACL

Eisenstein 2013. What to do about bad language on the internet. NAACL

Halliday, Matthiessen 2004. An introduction to functional grammar.

Han, Baldwin 2011. Lexical Normalisation of Short Text Messages: Makn Sens a# twitter. ACL

Hu, Talamadupula and Kambhampati 2013. Dude, srsly?: The Surprisingly Formal Nature of Twitter's Language. ICWSM

Ito, Tagliamonte 2003. Well weird, right dodgy, very strange, really cool. Language in Society 32(2):257

Jones 2010. The changing face of spelling on the internet.

Labov 1972. Sociolinguistic patterns.

Lui, Baldwin 2012. langid. py: An off-the-shelf language identification tool. ACL

Maynard 2014. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. LREC

O'Connor, Krieger, Ahn 2010. TweetMotif: Exploratory Search and Topic Summarization for Twitter. ICWSM

Owoputi, O'Connor, Dyer, Gimpel, Schneider, Smith 2013. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. NAACL

Plank, Hovy, Søgaard 2014. Learning POS taggers with inter-annotator agreement loss. EACL

Preotiuc-Pietro, Samangooei, Cohn, Gibbins, Naranjan 2012. Trendminer: An architecture for real time analysis of social media data. RAMSS

Ritter, Clark, Mausam, Etzioni 2011. Named Entity Recognition in Tweets: An Experimental Study. EMNLP

Tagliamonte, Denis 2008. Linguistic ruin? lol! instant messaging and teen language. American Speech 83(1):3

Turian, Ratinov, Bengio 2010. Word representations: a simple and general method for semi-supervised learning. ACL

Yates 1996. Oral and written linguistic aspects of computer conferencing. Pragmatics and beyond