Text Processing of Social Media II

Lexical Normalisation; User Geolocation; Twitter POS Tagging

Timothy Baldwin

Talk Outline

Characteristics of Text in Social Media

As we discussed in the first lecture, social media text contains varying levels of "noise" (lexical, syntactic and otherwise), e.g.:

- Tell ppl u luv them cuz 2morrow is truly not promised.
- SUPER BOWL SUNDAY!!! Enjoy yourselves!!! Sunday morning GOODIES R sent out! C U 2Nyt!
- Follow @OFA today for more coverage of the gun violence petition delivery to Congress. #NotBackingDown #EarlyFF

Why is Social Media Text "Bad"

- Eisenstein [2013] identified the following possible contributing factors to "badness" in social media text:
  - Lack of literacy?

Why is Social Media Text "Bad"

- Eisenstein [2013] identified the following possible contributing factors to "badness" in social media text:
  - Lack of literacy? *no* [Drouin and Davis, 2009]

Why is Social Media Text "Bad"

- Eisenstein [2013] identified the following possible contributing factors to "badness" in social media text:
  - Lack of literacy? *no* [Drouin and Davis, 2009]
  - Length restrictions?

Why is Social Media Text "Bad"

- Eisenstein [2013] identified the following possible contributing factors to "badness" in social media text:
  - Lack of literacy? *no* [Drouin and Davis, 2009]
  - Length restrictions? *not primarily* [Eisenstein, 2013]

Why is Social Media Text "Bad"

- Eisenstein [2013] identified the following possible contributing factors to "badness" in social media text:
  - Lack of literacy? *no* [Drouin and Davis, 2009]
  - Length restrictions? *not primarily* [Eisenstein, 2013]
  - Text input method-driven?

Why is Social Media Text "Bad"

- Eisenstein [2013] identified the following possible contributing factors to "badness" in social media text:
  - Lack of literacy? *no* [Drouin and Davis, 2009]
  - Length restrictions? *not primarily* [Eisenstein, 2013]
  - Text input method-driven? *to some degree, yes* [Gouws et al., 2011b]

Why is Social Media Text "Bad"

- Eisenstein [2013] identified the following possible contributing factors to "badness" in social media text:
  - Lack of literacy? *no* [Drouin and Davis, 2009]
  - Length restrictions? *not primarily* [Eisenstein, 2013]
  - Text input method-driven? *to some degree, yes* [Gouws et al., 2011b]
  - Pragmatics (mimicking prosodic effects etc. in speech)?

Why is Social Media Text "Bad"

- Eisenstein [2013] identified the following possible contributing factors to "badness" in social media text:
  - Lack of literacy? *no* [Drouin and Davis, 2009]
  - Length restrictions? *not primarily* [Eisenstein, 2013]
  - Text input method-driven? *to some degree, yes* [Gouws et al., 2011b]
  - Pragmatics (mimicking prosodic effects etc. in speech)? *yeeees* [Eisenstein, 2013]

Why is Social Media Text "Bad"

- Eisenstein [2013] identified the following possible contributing factors to "badness" in social media text:
  - Lack of literacy? *no* [Drouin and Davis, 2009]
  - Length restrictions? *not primarily* [Eisenstein, 2013]
  - Text input method-driven? *to some degree, yes* [Gouws et al., 2011b]
  - Pragmatics (mimicking prosodic effects etc. in speech)? *yeeees* [Eisenstein, 2013]
  - Social variables/markers of social identity?

Why is Social Media Text "Bad"

- Eisenstein [2013] identified the following possible contributing factors to "badness" in social media text:
  - Lack of literacy? *no* [Drouin and Davis, 2009]
  - Length restrictions? *not primarily* [Eisenstein, 2013]
  - Text input method-driven? *to some degree, yes* [Gouws et al., 2011b]
  - Pragmatics (mimicking prosodic effects etc. in speech)? *yeeees* [Eisenstein, 2013]
  - Social variables/markers of social identity? *blood oath!* [Eisenstein, 2013]

What can be done about it?

- **Lexical normalisation**: translate expressions into their canonical form
  - Issues:
    - What are the candidate tokens for normalisation?
    - To what degree do we allow normalisation?
    - What is the canonical form of a given expression? (e.g. *aint*)
    - Is lexical normalisation always appropriate? (e.g. *bro*, *pommie*)

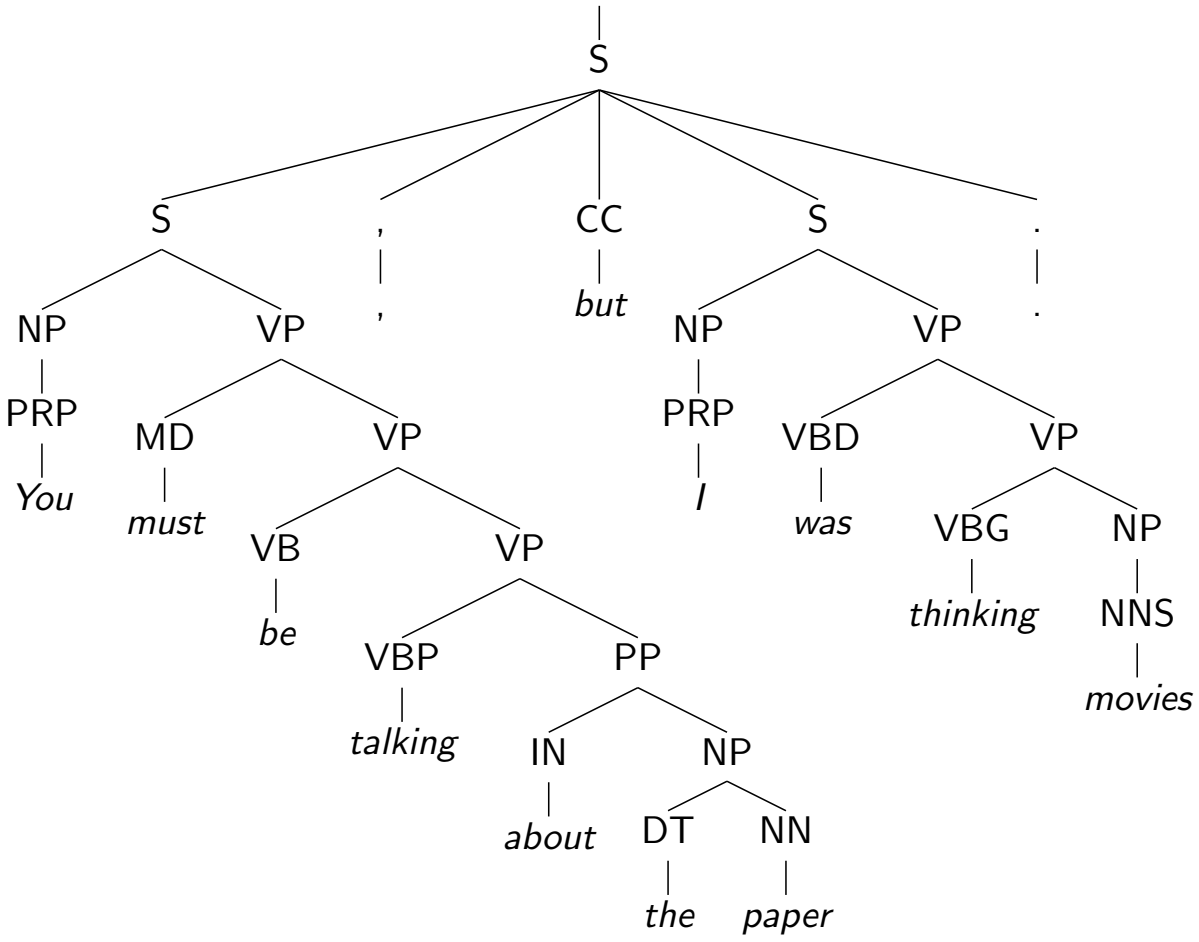
Source(s): Eisenstein [2013]

The Impact of Lexical Variation on NLP

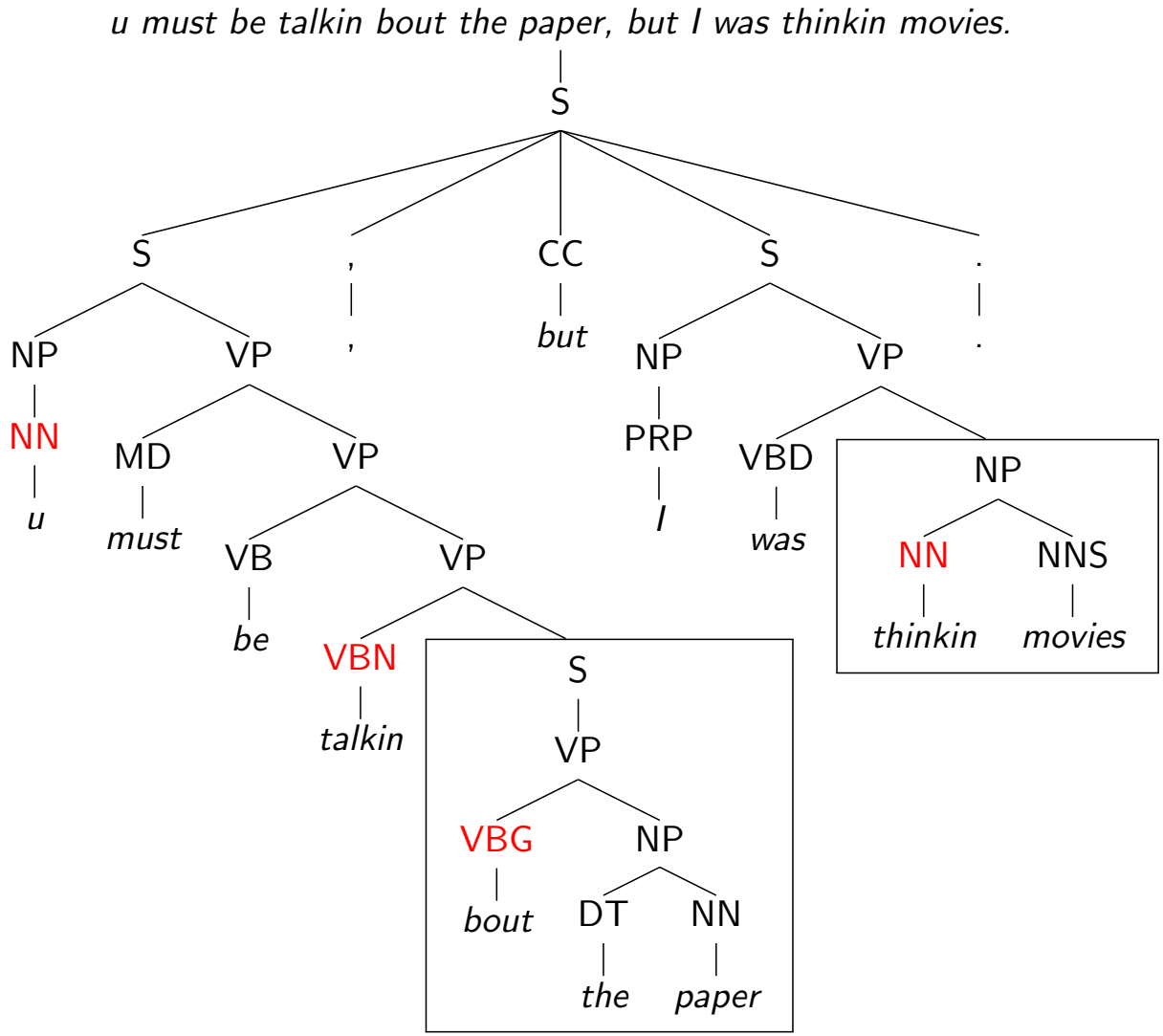
- We expect lexical and syntactic noise to negatively impact on content analysis when applied to social media text in (at least) the following ways:
  - loss of recall for keyword-based search
  - loss of accuracy of NLP technologies (POS taggers, parsers, NE recognisers, relation extractors, ...)

POS tagging and Parsing of Well-formed Text

You must be talking about the paper, but I was thinking movies.



POS tagging and Parsing of Noisy Twitter Text



Task Definition

- One standard definition of "lexical normalisation" is:
  - relative to some standard tokenisation
  - consider only OOV tokens as candidates for normalisation
  - allow only one-to-one (IV) word substitutions

i left ACL *cus* *im* *sickk* ! *Yuu* *better be their* *tmrw*

⇓ ⇓ ⇓ ⇓ ⇓

i left ACL *because* *I'm* *sick* ! *You* *better be their* *tomorrow*

Task Definition

- One standard definition of "lexical normalisation" is:
  - relative to some standard tokenisation
  - consider only OOV tokens as candidates for normalisation
  - allow only one-to-one (IV) word substitutions

i left ACL *cus* *im* *sickk* ! *Yuu better be their* *tmrw*

⇓ ⇓ ⇓ ⇓ ⇓

i left ACL *because* *I'm* *sick* ! *You better be their* *tomorrow*

- Assumptions/corrolaries of the task definition:
  - not possible to normalise IV tokens, e.g., *their* ("there")
  - not possible to normalise to multiword tokens, e.g., ignore *ttyl* ("talk to you later")
  - ignore Twitter-specific entities, e.g., @obama, #mandela, `bit.ly/1iRQeMe`
  - assumes a unique correct "norm" for each token

Categorisation of Non-standard Words in English Twitter

Category	**Ratio**	Example
Letter&Number	2.36%	*b4* "before"
Letter	72.44%	*shuld* "should"
Number Substitution	2.76%	*4* "for"
Slang	12.20%	*lol* "laugh out loud"
Other	10.24%	*sucha* "such a"

Table : Types of non-standard words in a 449 message sample of English tweets

Most non-standard words in sampled tweets are morphophonemic errors.

Source(s): Han and Baldwin [2011]

Categorisation of Non-standard Words in English Twitter

Category	**Ratio**	Example
Letter&Number	2.36%	*b4* "before"
Letter	72.44%	*shuld* "should"
Number Substitution	2.76%	*4* "for"
Slang	12.20%	*lol* "laugh out loud"
Other	10.24%	*sucha* "such a"

Table : Types of non-standard words in a 449 message sample of English tweets

Most non-standard words in sampled tweets are morphophonemic errors.

Source(s): Han and Baldwin [2011]

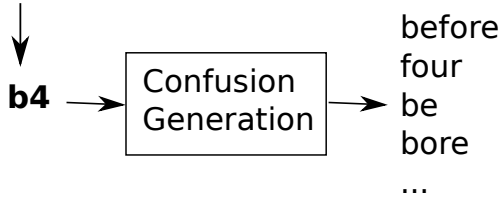
Token-based Approaches

1. Confusion set generation (i.e., find correction candidates)
2. Non-standard word detection (i.e., is the OOV a non-standard word?)
3. Normalisation of a non-standard word (i.e., select the candidate)

Token-based Approaches

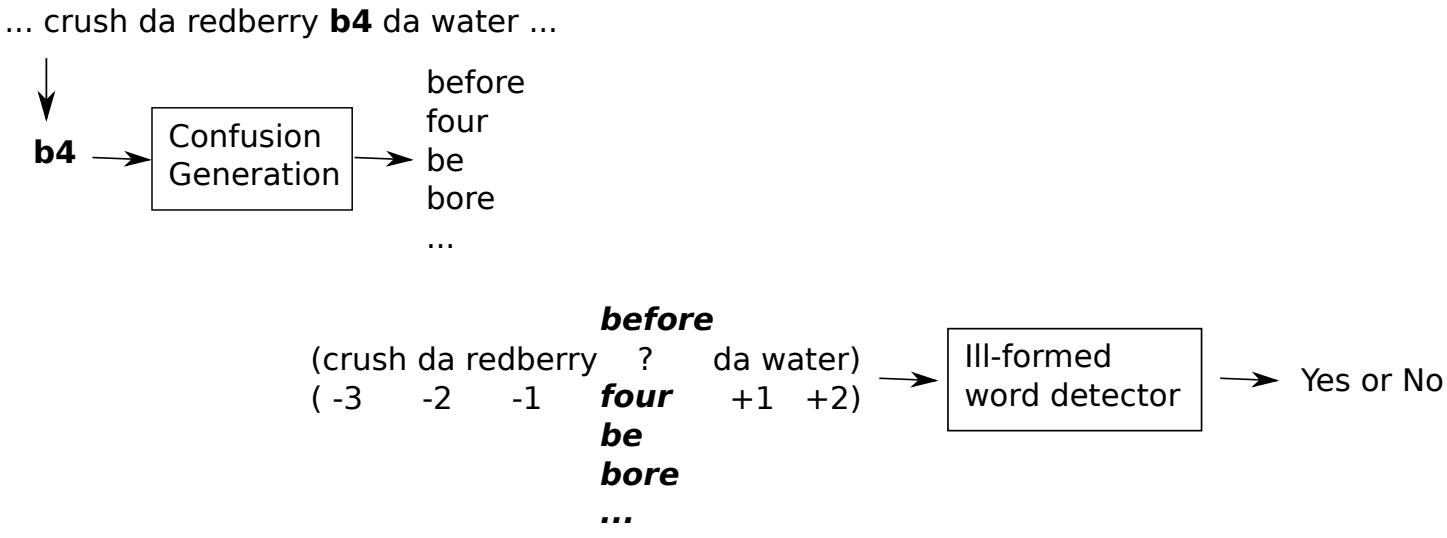
1. Confusion set generation (i.e., find correction candidates)
2. Non-standard word detection (i.e., is the OOV a non-standard word?)
3. Normalisation of a non-standard word (i.e., select the candidate)

... crush da redberry **b4** da water ...



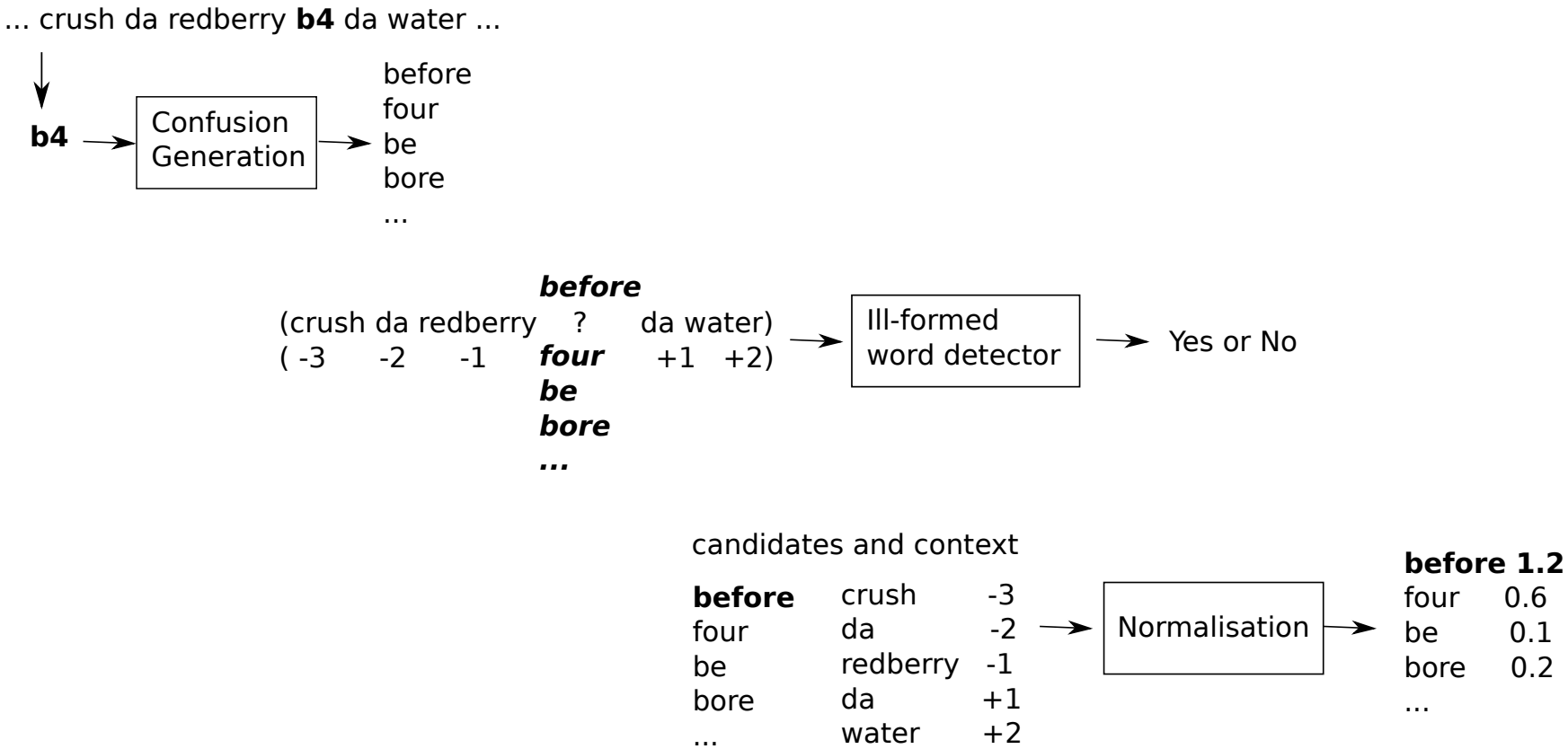
Token-based Approaches

1. Confusion set generation (i.e., find correction candidates)
2. Non-standard word detection (i.e., is the OOV a non-standard word?)
3. Normalisation of a non-standard word (i.e., select the candidate)



Token-based Approaches

1. Confusion set generation (i.e., find correction candidates)
2. Non-standard word detection (i.e., is the OOV a non-standard word?)
3. Normalisation of a non-standard word (i.e., select the candidate)



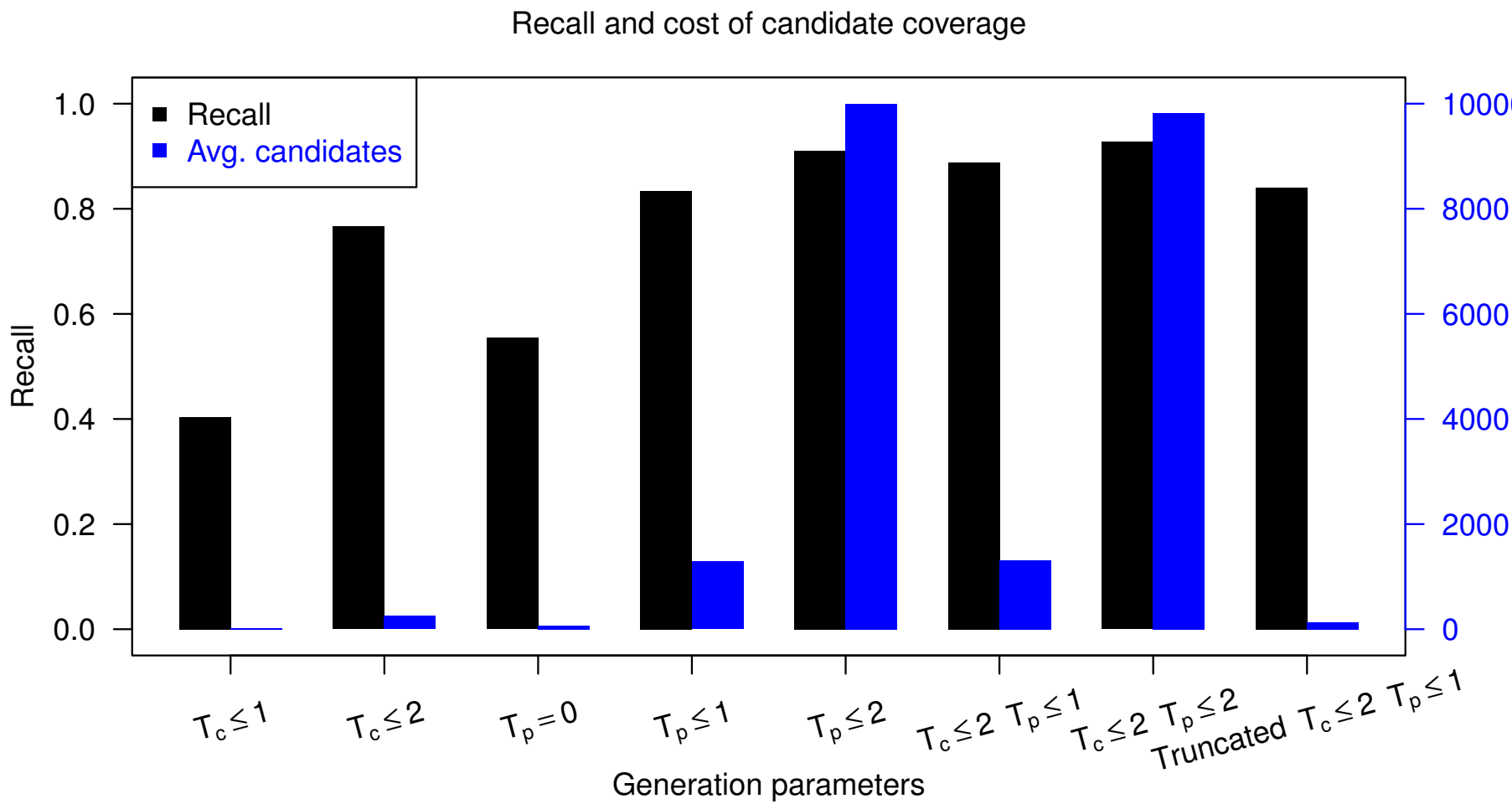
Pre-processing

- For OOV words, shorten any repetitions of 3+ letters to 2 letters
- Filter out any Twitter-specific tokens (user mentions, hashtags, `RT`, etc.) and URLs
- Identify all OOV words relative to a standard spelling dictionary (`aspell`)

Source(s): Han and Baldwin [2011]

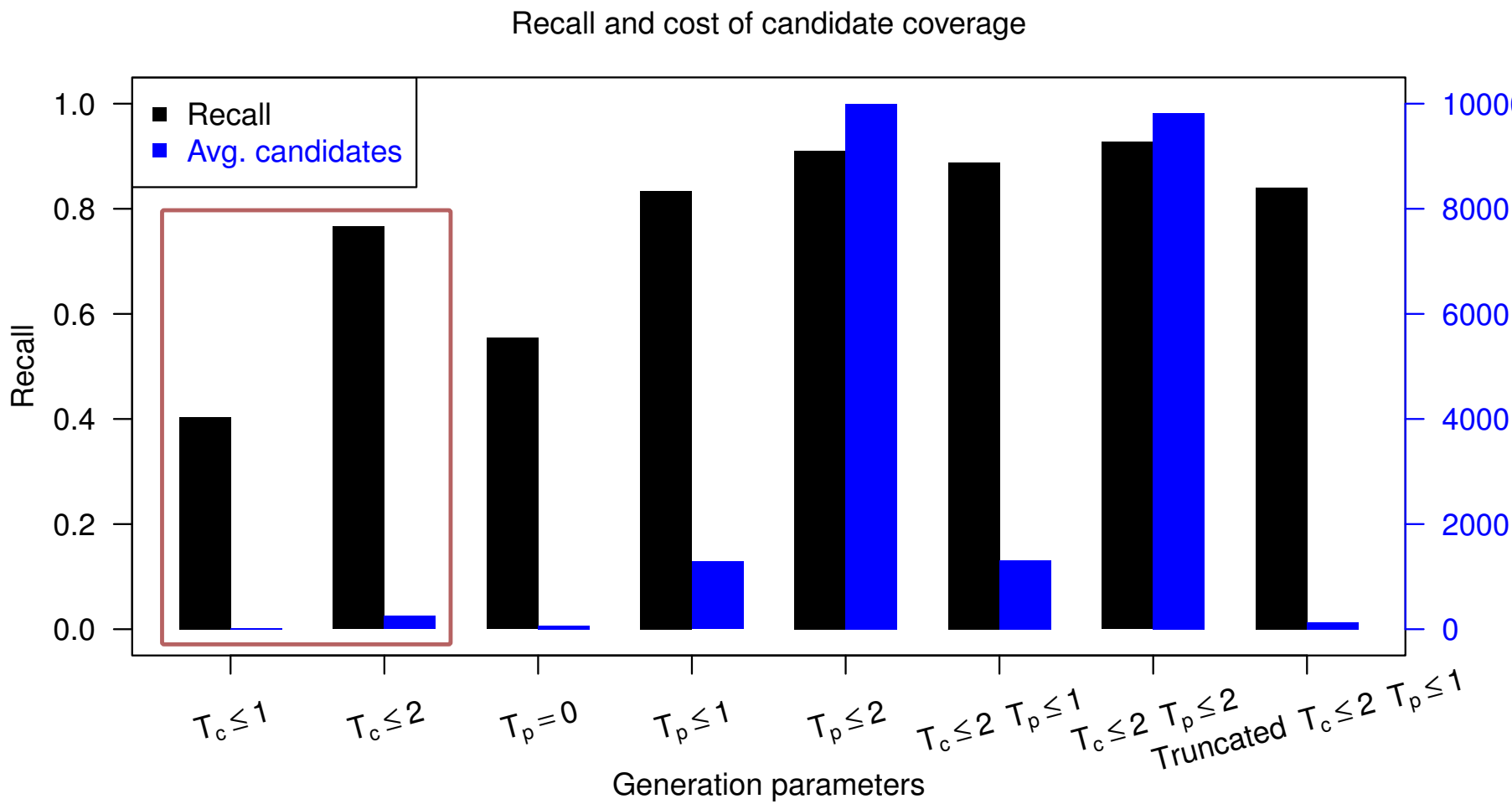
Candidate Generation

- Generation via edit distance over letters (T_c) and phonemes (T_p) [Philips, 2000], e.g. "love" and "laugh" are candidates for *luv* within $T_c \leq 2$ and $T_p = 0$.



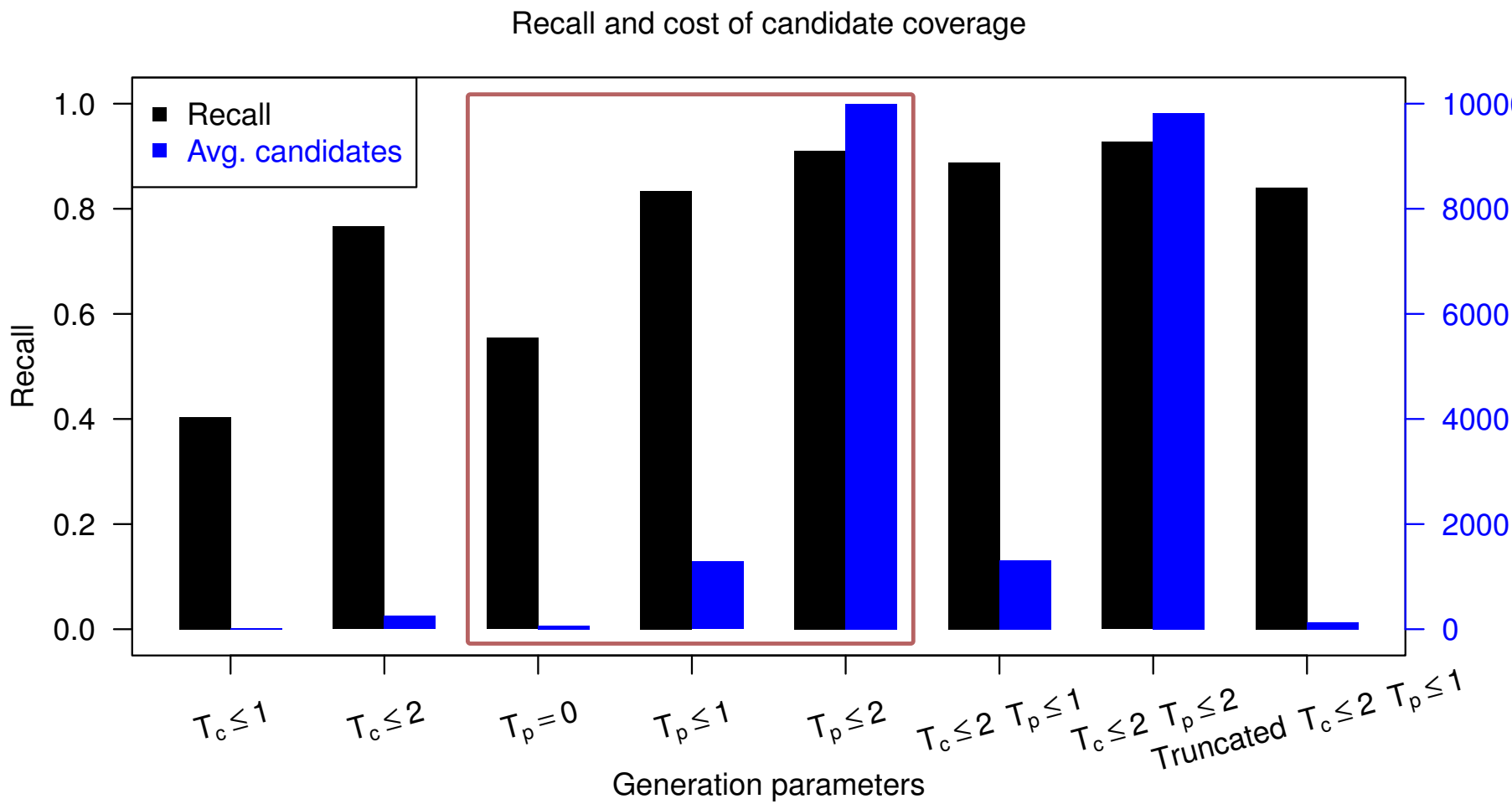
Candidate Generation

- Generation via edit distance over letters (T_c) and phonemes (T_p) [Philips, 2000], e.g. "love" and "laugh" are candidates for *luv* within $T_c \leq 2$ and $T_p = 0$.



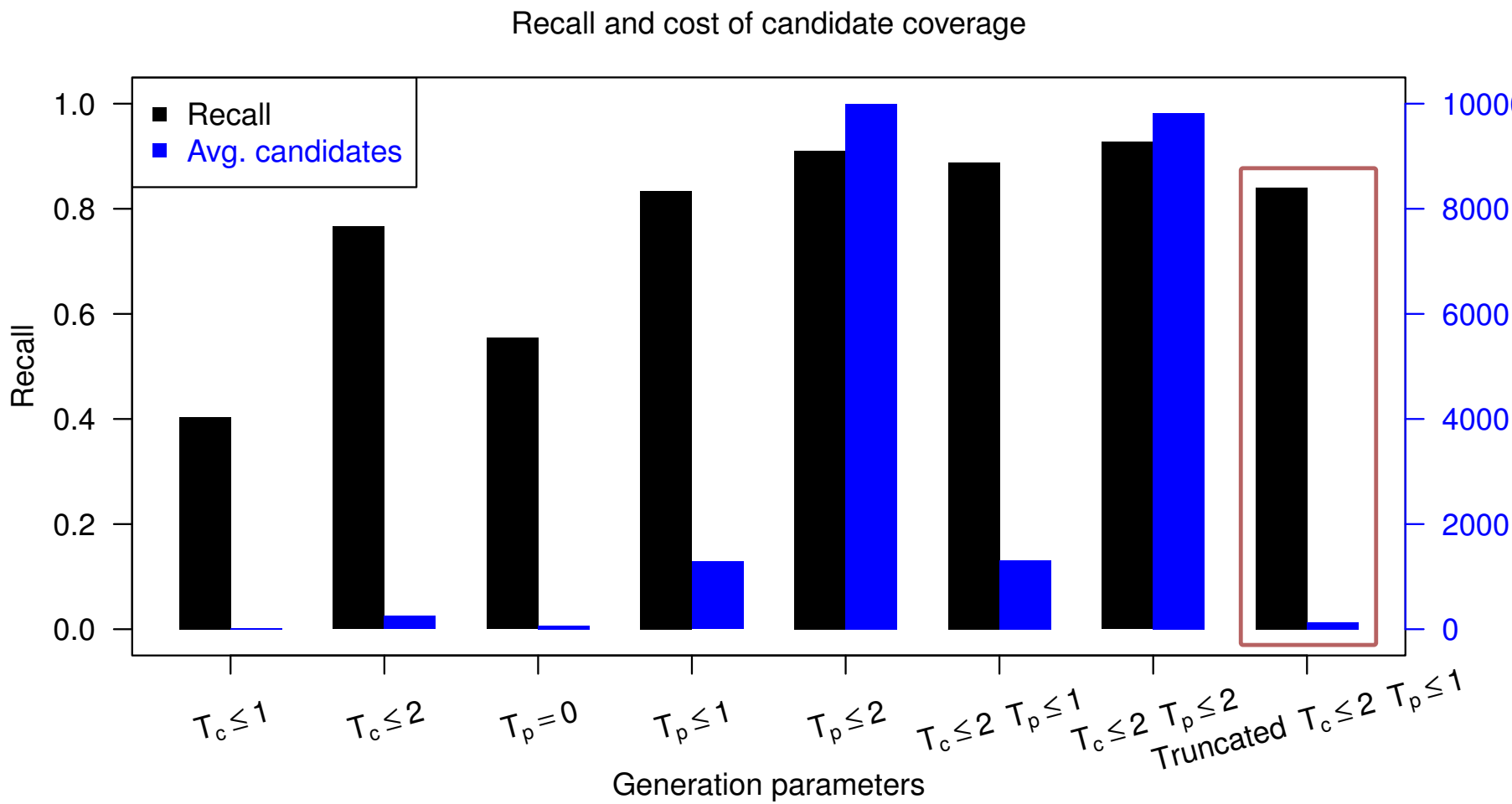
Candidate Generation

- Generation via edit distance over letters (T_c) and phonemes (T_p) [Philips, 2000], e.g. "love" and "laugh" are candidates for *luv* within $T_c \leq 2$ and $T_p = 0$.



Candidate Generation

- Generation via edit distance over letters (T_c) and phonemes (T_p) [Philips, 2000], e.g. "love" and "laugh" are candidates for *luv* within $T_c \leq 2$ and $T_p = 0$.



Detection of Ill-formed Words

- Detection based on candidate context fitness:
  - correct words should fit better with context than substitution candidates
  - incorrect words should fit *worse* that substitution candidates

Ill-formed word in text snippet	**Candidate**	**Dependencies**
but I was thinkin movies .	(thinking, ...)	dobj(thinking, movies)
article poster by ruderrobb : there was	(rattrap, ...)	–

- A SVM classifier is trained based on dependencies, to indicate candidate context fitness.
  - How to generate features from dependencies?
  - How to train context fitness SVM classifier?
  - How to detect ill-formed word based on SVM outputs?

Source(s): Han and Baldwin [2011]

Feature Representation Using Dependencies

- Build a dependency bank from existing corpora
- Represent each dependency tuple as a word pair + positional index (*without the dependency type*)

Source(s): Han and Baldwin [2011]

Feature Representation Using Dependencies

- Build a dependency bank from existing corpora
- Represent each dependency tuple as a word pair + positional index (*without the dependency type*)

Corpus (NYT)

One obvious difference is the way they look, ...

↓ **Stanford Parser**

num(difference-3, One-1)
amod(difference-3, obvious-2)
nsubj(way-6, difference-3)
cop(way-6, is-4)
det(way-6, the-5)
dobj(look-8, way-6)
nsubj(look-8, they-7)
rcmod(way-6, look-8)
...

→

Dependency bank

(way, difference, 3)
(look, way, 2)
...

Source(s): Han and Baldwin [2011]

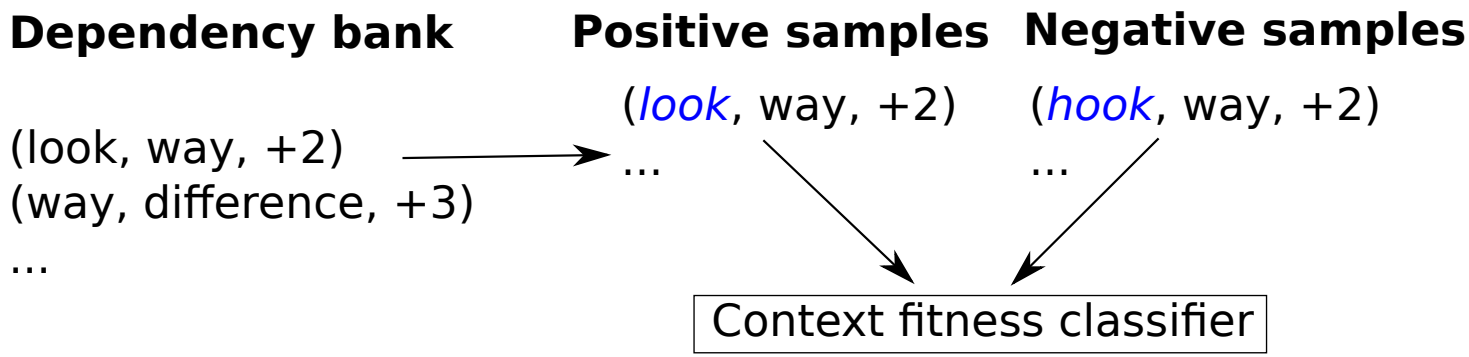
SVM Training Data Generation

- Use dependency bank directly as positive features
- Automatically generate negative dependency features by replacing the target word with highly-ranked confusion candidates

Source(s): Han and Baldwin [2011]

SVM Training Data Generation

- Use dependency bank directly as positive features
- Automatically generate negative dependency features by replacing the target word with highly-ranked confusion candidates



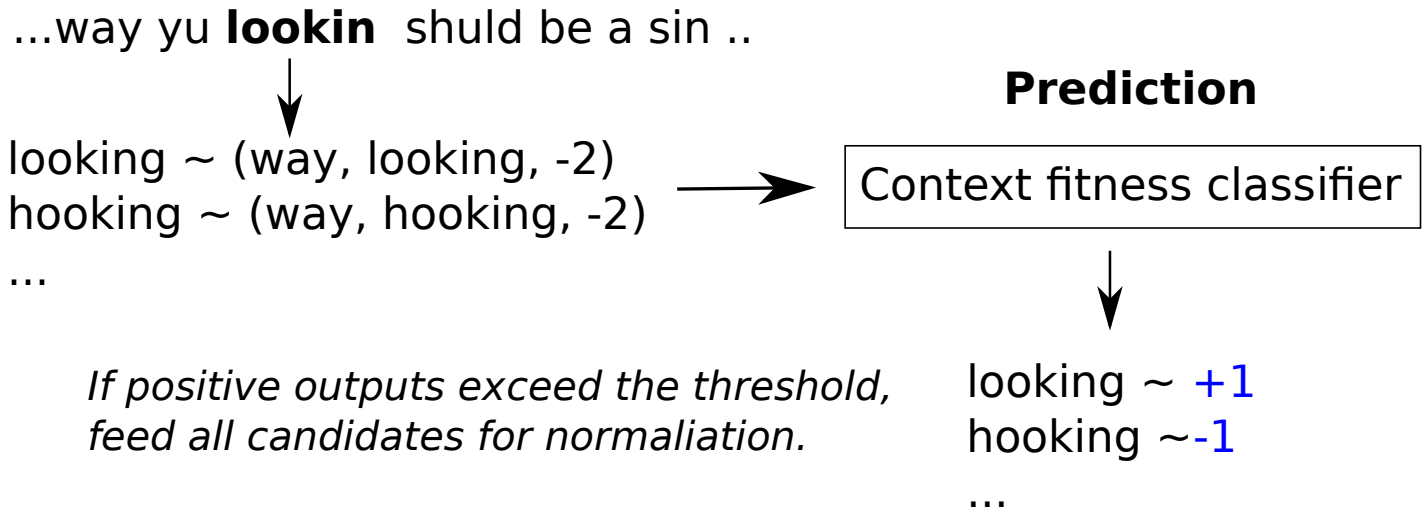
Source(s): Han and Baldwin [2011]

Putting it together: Detecting Ill-formed Words

- OOV words with candidates fitting the context (i.e. positive classification outputs) are probably ill-formed words

Putting it together: Detecting Ill-formed Words

- OOV words with candidates fitting the context (i.e. positive classification outputs) are probably ill-formed words



- Set threshold=1 $\Rightarrow$ *lookin* is considered to be an ill-formed word

Source(s): Han and Baldwin [2011]

Normalisation Candidate Selection

- For each ill-formed word and its possible correction candidates, the following features are consider for normalisation:
  - Word Similarity (WS)
    - letter and phoneme edit distance (ED)
    - prefix, suffix, and longest common subsequence

Normalisation Candidate Selection

- For each ill-formed word and its possible correction candidates, the following features are consider for normalisation:
  - Word Similarity (WS)
    - letter and phoneme edit distance (ED)
    - prefix, suffix, and longest common subsequence
  - Context Support (CS)
    - trigram language model score
    - dependency score (weighted dependency count, derived from the detection step)

Source(s): Han and Baldwin [2011]

Other Token-based Approaches

- SMT over gold-standard training data [Aw et al., 2006], or silver-standard training data generated through self-translation detection + MT [Ling et al., 2013]
- Unsupervised log-linear sequential labelling model, trained over a set of normalisation candidates and language model [Yang and Eisenstein, 2013]
- Random walk models applies to normalisation candidate sets, optimised relative to a language model [Zhang et al., 2013]

Type-based Approach: Outline

- **Observation:** the longer the ill-formed word, the more likely there is a unique normalisation candidate:
  - *y* $\Rightarrow$ {*why*, *you*, . . . }, *hw* $\Rightarrow$ {*how*, *homework*, . . . }
  - *4eva* $\Rightarrow$ {forever}, *tlkin* $\Rightarrow$ {*talking*}
- **Approach:** construct a dictionary of (lexical variant, standard form) pairs for longer word types (character length ≥ 4) of moderate frequency (≥ 16)

Source(s): Han et al. [2012a]

Type-based Approach: Approach

- Construct the dictionary based on distributional similarity + string similarity:

  Input: Tokenised English tweets

  1. Extract (OOV, IV) pairs based on distributional similarity.
  2. Re-rank the extracted pairs by string similarity.

  Output: A list of (OOV, IV) pairs ordered by string similarity; select the top-n pairs for inclusion in the normalisation lexicon.

Source(s): Han et al. [2012a]

An Example

... see you tmrw ...
... tmrw morning ...
... tomorrow morning ...
...

$\Downarrow$ distributional similarity

$\{$*tmrw, 2morow, tomorrow, Monday*$\}$

$\Downarrow$ string similarity

tmrw $\rightarrow$ *tomorrow*

Source(s): Han et al. [2012a]

Context Modelling

- Components/parameters of the method:
  - context window size: ± 1, ± 2, ± 3
  - context word sensitivity: bag-of-words vs. positional indexing
  - context word representation: unigram, bigram or trigram
  - context word filtering: all tokens vs. only dictionary words
  - context similarity: KL divergence, Jensen-Shannon divergence, Cosine similarity, Euclidean distance
- Tune parameters relative to (OOV, IV) pair development data

Source(s): Han et al. [2012a]

Context Modelling

- Components/parameters of the method:
  - context window size: ± 1, ± 2, ± 3
  - context word sensitivity: bag-of-words vs. positional indexing
  - context word representation: unigram, bigram or trigram
  - context word filtering: all tokens vs. only dictionary words
  - context similarity: KL divergence, Jensen-Shannon divergence, Cosine similarity, Euclidean distance
- Tune parameters relative to (OOV, IV) pair development data

Source(s): Han et al. [2012a]

Rerank Pairs By String Similarity

- (OOV, IV) pairs derived by distributional similarity:

  (*Obama*, *Adam*)
  (*tmrw*, *tomorrow*)
  (*Youtube*, *web*)
  (*4eva*, *forever*)
  ...

Source(s): Han et al. [2012a]

Rerank Pairs By String Similarity

- (OOV, IV) pairs derived by distributional similarity:

  (*Obama*, *Adam*) ↓
  (*tmrw*, *tomorrow*) ↑
  (*Youtube*, *web*) ↓
  (*4eva*, *forever*) ↑
  ...

Source(s): Han et al. [2012a]

Rerank Pairs By String Similarity

- (OOV, IV) pairs derived by distributional similarity:

  (*Obama*, *Adam*) ↓
  (*tmrw*, *tomorrow*) ↑
  (*Youtube*, *web*) ↓
  (*4eva*, *forever*) ↑
  ...

- Get the top-ranked pairs as lexicon entries:

  (*tmrw*, *tomorrow*)
  (*4eva*, *forever*)
  (*Obama*, *Adam*)
  (*Youtube*, *web*)
  ...

Source(s): Han et al. [2012a]

Other Type-based Approaches

- Noisy Channel approach:

$$\arg\max P(t|s) = \arg\max_t P(s|t)P(t)$$

  Model $P(t)$ using language model or HMM; model $P(s|t)$ using some error model [Brill and Moore, 2000, Choudhury et al., 2007, Cook and Stevenson, 2009]
- Source silver-standard type-level training data using Google query correction, and use as the basis for training a character-level CRF [Liu et al., 2011]
- Automatically construct a normalisation lexicon through distributional similarity + dictionary filtering + lexical similarity [Gouws et al., 2011a]

Evaluation I

- **Lexnorm dataset:** 549 pre-tokenised messages with 1184 non-standard words, and pre-identification of OOV words, e.g.:

```
new      IV     new
pix      OOV    pictures
comming  OOV    coming
tomoroe  OOV    tomorrow
```

- Evaluation in terms of:
  - OOV token-level precision, recall and F-score
  - word error rate (WER) over all tokens
  - false alarm rate (FA) over normalised tokens tokens
  - BLEU (a la MT)
- NB many papers on lexical normalisation assume oracle pre-determination of ill-formed words(!)

Evaluation II

- For example:

  new *pix* *comming* *tomoroe*

  ⇓ ⇓ ⇓

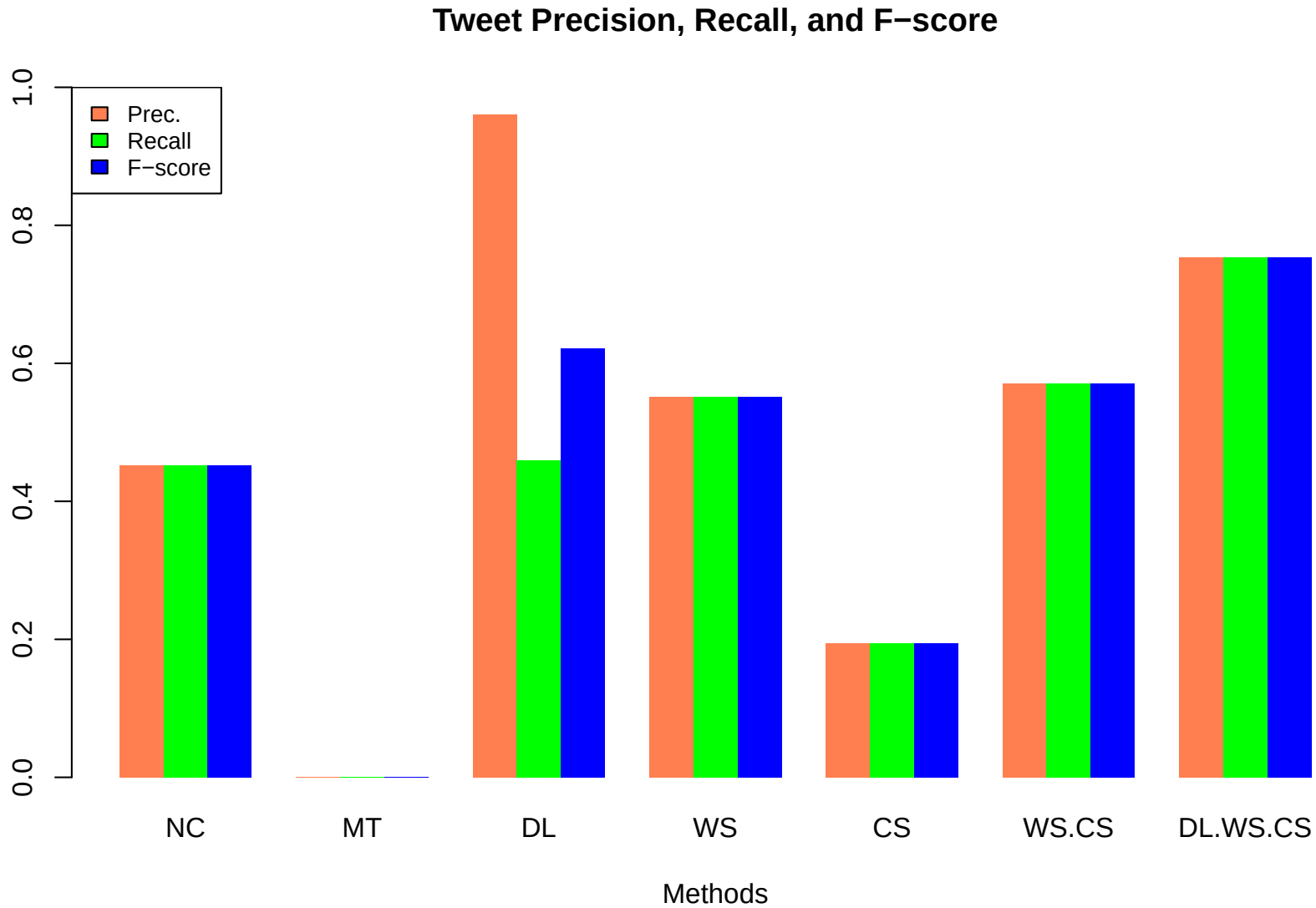
  picks *coming* —

  $\mathcal{P} = \frac{1}{2}$, $\mathcal{R} = \frac{1}{3}$, $\mathcal{F} = 0.4$

  WER $= \frac{2}{4}$

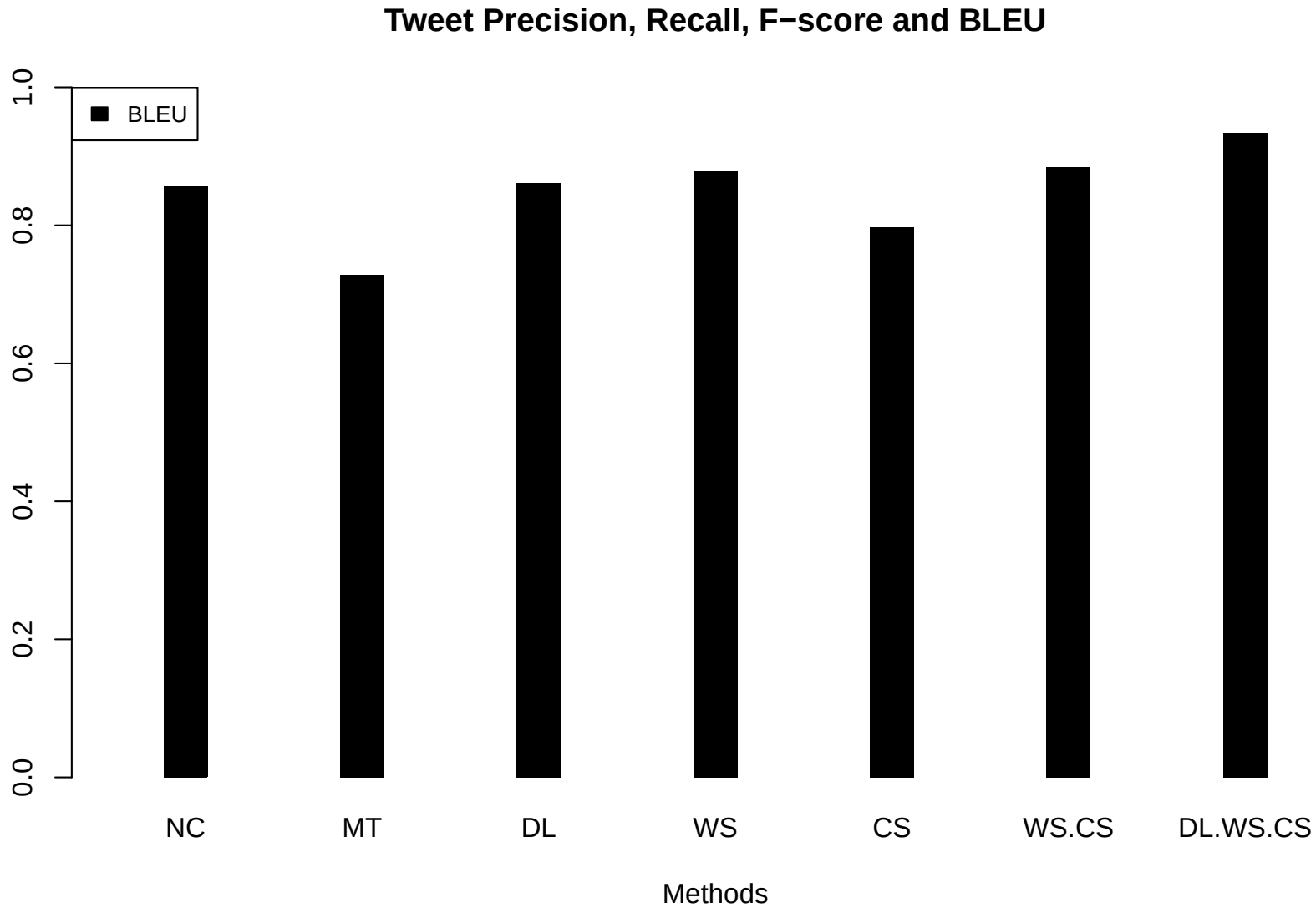
  FA $= \frac{2}{4}$

Intrinsic Evaluation (Token-based Method)



NC = noisy channel model; *MT* = phrasal machine translation;
DL = dictionary lookup; *WS* = word similarity; *CS* = context support

Intrinsic Evaluation (Token-based Method)



NC = noisy channel model; *MT* = phrasal machine translation;
DL = dictionary lookup; *WS* = word similarity; *CS* = context support

Intrinsic Evaluation (Type-based Method)

Method	Precision	Recall	F-Score
1. Internet slang dictionary	0.915	0.435	0.590
2. Gouws et al. [2011a] lexicon	**0.982**	0.319	0.482
3. Our proposed lexicon	0.700	0.179	0.285
1. + 3.	0.840	0.601	0.701
2. + 3.	0.863	0.498	0.632
1. + 2.	0.920	0.465	0.618
1. + 2. + 3.	0.847	0.630	**0.723**
End-to-end token-based method	0.515	**0.771**	0.618

Table : Individual and Combined Lexicon-based Normalisation Results

Intrinsic Evaluation (Type-based Method)

Method	Precision	Recall	F-Score
1. Internet slang dictionary	0.915	0.435	0.590
2. Gouws et al. [2011a] lexicon	**0.982**	0.319	0.482
3. Our proposed lexicon	0.700	0.179	0.285
1. + 3.	0.840	0.601	0.701
2. + 3.	0.863	0.498	0.632
1. + 2.	0.920	0.465	0.618
1. + 2. + 3.	0.847	0.630	**0.723**
End-to-end token-based method	0.515	**0.771**	0.618

Table : Individual and Combined Lexicon-based Normalisation Results

- Derived lexicon covers the long tail of non-standard words and is complementary to existing slang lexicons.
- Best end-to-end normalisation system at the time

Intrinsic Evaluation (Type-based Method)

Method	Precision	Recall	F-Score
1. Internet slang dictionary	0.915	0.435	0.590
2. Gouws et al. [2011a] lexicon	**0.982**	0.319	0.482
3. Our proposed lexicon	0.700	0.179	0.285
1. + 3.	0.840	0.601	0.701
2. + 3.	0.863	0.498	0.632
1. + 2.	0.920	0.465	0.618
1. + 2. + 3.	0.847	0.630	**0.723**
End-to-end token-based method	0.515	**0.771**	0.618

Table : Individual and Combined Lexicon-based Normalisation Results

- Derived lexicon covers the long tail of non-standard words and is complementary to existing slang lexicons.
- Best end-to-end normalisation system at the time

Intrinsic Evaluation (Type-based Method)

Method	Precision	Recall	F-Score
1. Internet slang dictionary	0.915	0.435	0.590
2. Gouws et al. [2011a] lexicon	**0.982**	0.319	0.482
3. Our proposed lexicon	0.700	0.179	0.285
1. + 3.	0.840	0.601	0.701
2. + 3.	0.863	0.498	0.632
1. + 2.	0.920	0.465	0.618
1. + 2. + 3.	0.847	0.630	**0.723**
End-to-end token-based method	0.515	**0.771**	0.618

Table : Individual and Combined Lexicon-based Normalisation Results

- Derived lexicon covers the long tail of non-standard words and is complementary to existing slang lexicons.
- Best end-to-end normalisation system at the time

Intrinsic Evaluation (Type-based Method)

Method	Precision	Recall	F-Score
1. Internet slang dictionary	0.915	0.435	0.590
2. Gouws et al. [2011a] lexicon	**0.982**	0.319	0.482
3. Our proposed lexicon	0.700	0.179	0.285
1. + 3.	0.840	0.601	0.701
2. + 3.	0.863	0.498	0.632
1. + 2.	0.920	0.465	0.618
1. + 2. + 3.	0.847	0.630	**0.723**
End-to-end token-based method	0.515	**0.771**	0.618

Table : Individual and Combined Lexicon-based Normalisation Results

- Derived lexicon covers the long tail of non-standard words and is complementary to existing slang lexicons.
- Best end-to-end normalisation system at the time

Extrinsic Evaluation (Type-based Method)

- **Task:** analyse the impact of text normalisation on Twitter POS tagging
- **Dataset:** annotated tweets from CMU ARK Twitter POS tagger test data [Gimpel et al., 2011b]

Tagger	Text	% accuracy
Stanford POS tagger	original	68.4
Stanford POS tagger	normalised	70.0
Twitter POS tagger	original	95.2
Twitter POS tagger	normalised	94.7

Extrinsic Evaluation (Type-based Method)

- **Task:** analyse the impact of text normalisation on Twitter POS tagging
- **Dataset:** annotated tweets from CMU ARK Twitter POS tagger test data [Gimpel et al., 2011b]

Tagger	Text	% accuracy
Stanford POS tagger	original	68.4
Stanford POS tagger	normalised	70.0
Twitter POS tagger	original	95.2
Twitter POS tagger	normalised	94.7

- Text normalisation improves POS tagging accuracy

Extrinsic Evaluation (Type-based Method)

- **Task:** analyse the impact of text normalisation on Twitter POS tagging
- **Dataset:** annotated tweets from CMU ARK Twitter POS tagger test data [Gimpel et al., 2011b]

Tagger	Text	% accuracy
Stanford POS tagger	original	68.4
Stanford POS tagger	normalised	70.0
Twitter POS tagger	original	95.2
Twitter POS tagger	normalised	94.7

- Text normalisation improves POS tagging accuracy
- Also successful applied to NER, MT and event detection

Spanish Text Normalisation Task

- **Opportunities:**
  - orthography: Latin script
  - word segmentation: whitespace
  - morphophonemics: similar to English
  - lexical resources: Internet slang dictionary
- **Challenges:**
  - capitalisation: case-sensitive
  - evaluation: accuracy
  - scope: one-to-many normalisation, e.g., *tqm* ("te quiero mucho")
- **Data:** 564 annotated Spanish tweets from tweet-norm shared task

Spanish Text Normalisation Task

- **Opportunities:**
  - orthography: Latin script
  - word segmentation: whitespace
  - morphophonemics: similar to English
  - lexical resources: Internet slang dictionary
- **Challenges:**
  - capitalisation: case-sensitive
  - evaluation: accuracy
  - scope: one-to-many normalisation, e.g., *tqm* ("te quiero mucho")
- **Data:** 564 annotated Spanish tweets from tweet-norm shared task

Adapt English lexicon-based method to Spanish

Spanish Text Normalisation Results

- Accuracy of lexicon-based normalisation system ("–" indicates the ablation of a particular lexicon):

Lexicon	Accuracy
Combined lexicon	0.52
–slang lexicon	0.51
–dev lexicon	0.46
–dist lexicon	0.42
–name lexicon	0.51
Baseline	0.20

- Automatically-derived lexicon ("dist lexicon") is complementary to existing lexicons.

Spanish Text Normalisation Results

- Accuracy of lexicon-based normalisation system ("–" indicates the ablation of a particular lexicon):

Lexicon	Accuracy
Combined lexicon	0.52
–slang lexicon	0.51
–dev lexicon	0.46
–dist lexicon	0.42
–name lexicon	0.51
Baseline	0.20

- Automatically-derived lexicon ("dist lexicon") is complementary to existing lexicons.

The Road Ahead

- One-to-many and many-to-many normalisation still largely out of reach of current methods
- Better (but efficient) context sensitisation of lexical normalisation methods
- Better standardisation of evaluation (incl. more realistic evaluation setup)
- Interesting work to be done investigating the causes of lexical variation, and further investigating the question of what is "normal"
- Interactions between lexical normalisation and user profiling

Summary of Work on Lexical Normalisation

- Compare existing methods on tweet text normlaisation
- A token-based method using morphonphomeic clues
- A type-based method using distributional and string similarities
- Text normalisation improves downstream POS tagging
- Normalisation lexicons are lightweight, easy to use and well generalised to other languages (e.g., Spanish).

Talk Outline

User Geolocations

```
USER1: hinsdale-il043-us
USER2: fresno-ca019-us
USER3: el monte-ca037-us
USER4: elkridge-md027-us
USER5: owensboro-ky059-us
USER6: washington, d.c.-dc001-us
USER7: pasadena-ca037-us
USER8: daly city-ca081-us
USER9: youngstown-oh099-us
USER10: redan-ga089-us
```

Why Geolocation Information?

Event detection

Sentiment analysis

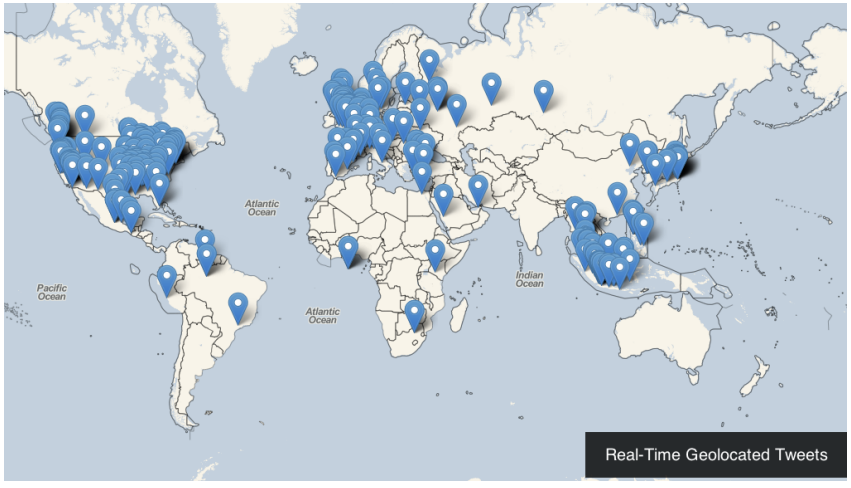
Advertising

Recommendation

Research Motivation: Geospatial Information

Event detection

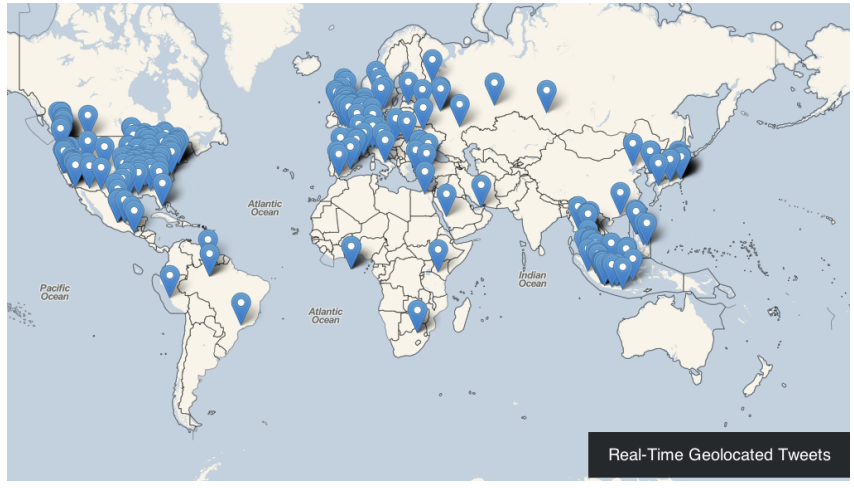
?



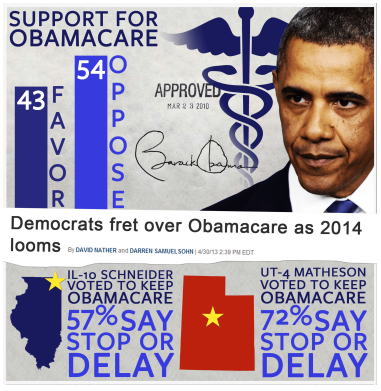
Research Motivation: Geospatial Information

Event detection

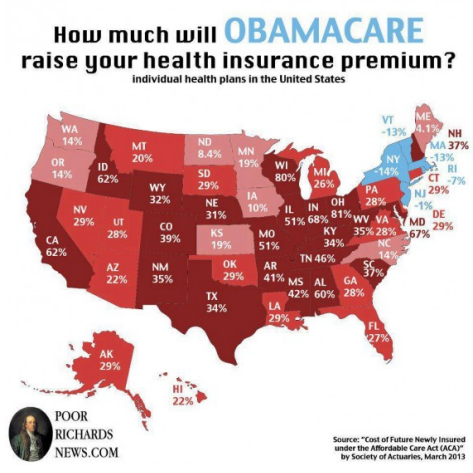
?



Sentiment analysis



?



Predict Geolocations for Twitter Users

Challenges:

- IP-based method?

Predict Geolocations for Twitter Users

Challenges:

- IP-based method? Data availability issues

Predict Geolocations for Twitter Users

Challenges:

- IP-based method? Data availability issues
- Tweets with GPS labels (i.e., geotagged tweets)?

Predict Geolocations for Twitter Users

Challenges:

- IP-based method? Data availability issues
- Tweets with GPS labels (i.e., geotagged tweets)? 1–1.7%

Predict Geolocations for Twitter Users

Challenges:

- IP-based method? Data availability issues
- Tweets with GPS labels (i.e., geotagged tweets)? 1–1.7%
- User self-declared locations?

Predict Geolocations for Twitter Users

Challenges:

- IP-based method? Data availability issues
- Tweets with GPS labels (i.e., geotagged tweets)? 1–1.7%
- User self-declared locations? e.g., *in your heart*

Predict Geolocations for Twitter Users

Challenges:

- IP-based method? Data availability issues
- Tweets with GPS labels (i.e., geotagged tweets)? 1–1.7%
- User self-declared locations? e.g., *in your heart*

Research objective:

Predict a user's primary location based on the text content associated with their recent posts

Predict Geolocations for Twitter Users

Challenges:

- IP-based method? Data availability issues
- Tweets with GPS labels (i.e., geotagged tweets)? 1–1.7%
- User self-declared locations? e.g., *in your heart*

Research objective:

Predict a user's primary location based on the text content associated with their recent posts

Input: Aggregated tweets from a Twitter user $\Rightarrow$ **Output:** A location in the world

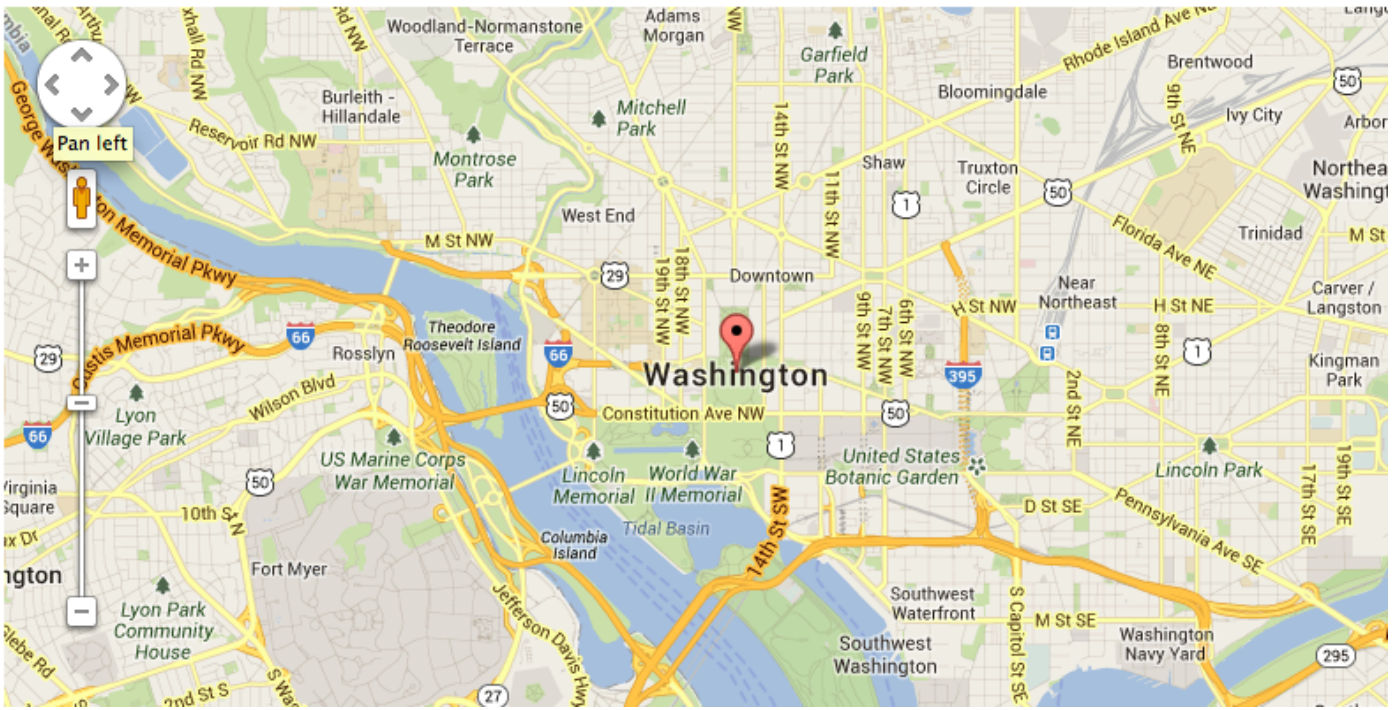
An Intuitive Example

Where is @BarackObama?

Please enter a user screen name, e.g. BarackObama, BBCNews. (Note: Only Google Chrome browser is supported.)

BarackObama Geolocate

- Prediction for **BarackObama: Washington, D. C., United States. (Latitude: 38.89511, Longitude: -77.03637)**
- Summary: **BarackObama** has **200** recent status updates. None of them is geotagged.



Source(s): `http://maps.google.com`, `http://hum.csse.unimelb.edu.au:9000/geo.html`

Can you Guess the City?

Geographic Representation of Location

- The main approaches to geographically representing user locations have been:
  - point-based

Geographic Representation of Location

- The main approaches to geographically representing user locations have been:
  - point-based
  - geopolitical (e.g. one of the 48 contiguous US states) [Eisenstein et al., 2010]

Geographic Representation of Location

- The main approaches to geographically representing user locations have been:
  - point-based
  - geopolitical (e.g. one of the 48 contiguous US states) [Eisenstein et al., 2010]
  - city-based (e.g. one of 3709 cities of a certain size in the world) [Han et al., 2012b]

Geographic Representation of Location

- The main approaches to geographically representing user locations have been:
  - point-based
  - geopolitical (e.g. one of the 48 contiguous US states) [Eisenstein et al., 2010]
  - city-based (e.g. one of 3709 cities of a certain size in the world) [Han et al., 2012b]
  - grid cell-based

Geographic Representation of Location

- The main approaches to geographically representing user locations have been:
  - point-based
  - geopolitical (e.g. one of the 48 contiguous US states) [Eisenstein et al., 2010]
  - city-based (e.g. one of 3709 cities of a certain size in the world) [Han et al., 2012b]
  - grid cell-based
    - latlong uniform-sized grid cell

Geographic Representation of Location

- The main approaches to geographically representing user locations have been:
  - point-based
  - geopolitical (e.g. one of the 48 contiguous US states) [Eisenstein et al., 2010]
  - city-based (e.g. one of 3709 cities of a certain size in the world) [Han et al., 2012b]
  - grid cell-based
    - latlong uniform-sized grid cell
    - population-based uniform-sized grid cell Roller et al. [2012]

Stochastic Representation of Location

- Additional dimension of how to stochastically represent the user:

Stochastic Representation of Location

- Additional dimension of how to stochastically represent the user:
  - "one-hot" (the user is at a unique location)

Stochastic Representation of Location

- Additional dimension of how to stochastically represent the user:
  - "one-hot" (the user is at a unique location)
  - discrete probabilistic (probability distribution over a discrete geographical representation)

Stochastic Representation of Location

- Additional dimension of how to stochastically represent the user:
  - "one-hot" (the user is at a unique location)
  - discrete probabilistic (probability distribution over a discrete geographical representation)
  - continuous probabilistic (2D probability density function over a region, based on a discrete geographical representation) [Priedhorsky et al., 2014]

Different Dimensions of Locative Aboutness

- Location can be described in terms of:
  - the *about* location
    - what location is a tweet about
  - the *tweeting* location
    - where was the tweet sent from (the GPS coordinates of the sending location)
  - a user's *primary* location
    - what is the "home base" of the user
- For example (sent from FRA, e.g.):

  En route to Saarland; looking forward to a productive stay

  about = Saarland, DE
  tweeting = Frankfurt, DE
  primary = Melbourne, AU

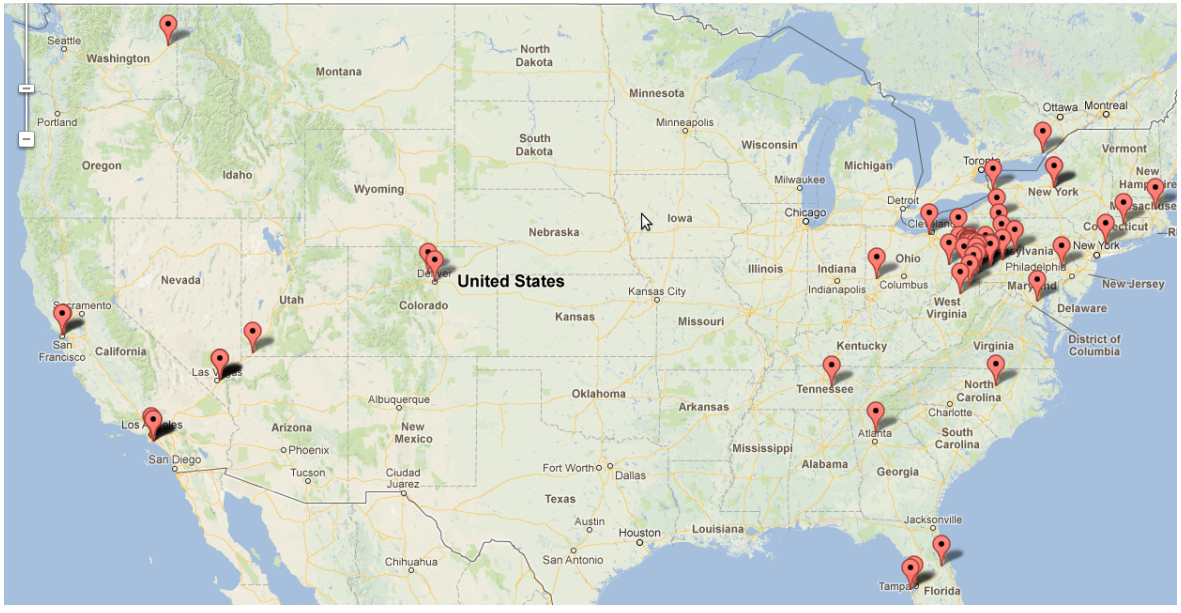
Source(s): Han et al. [2014]

Location Indicative Words

- "Location indicative words" (LIWs) either explicitly or implicitly infer locations, and take the form, e.g., of:
  - Gazetted terms: *Melbourne, London, Boston*
  - Dialectal terms: *tata, aloha*
  - Local features: *tram, tube, the G*
  - Sports terms: *footy, superbowl, hockey*
  - Weather related words: *cold, windy*

Varying Levels of Location Indicativeness

Local words : *yinz*, *hoagie*, *dippy*



Varying Levels of Location Indicativeness

Local words : *yinz*, *hoagie*, *dippy*

Somewhat local words : *ferry*, *chinatown*, *tram*

Varying Levels of Location Indicativeness

Local words : *yinz*, *hoagie*, *dippy*

Somewhat local words : *ferry*, *chinatown*, *tram*

Common words : *today*, *twitter*, *iphone*

Varying Levels of Location Indicativeness

Local words : *yinz*, *hoagie*, *dippy*

Somewhat local words : *ferry*, *chinatown*, *tram*

Common words : *today*, *twitter*, *iphone*

- How to identify and select location indicative words?
- Do LIWs improve prediction accuracy over using all tokens?

Feature Selection I

- LIW identification = feature selection, which can take a number of forms:
- Statistical:
  - chi square (CHI and MaxCHI)
  - log likelihood (LOGLIKE)
- Information-theoretic:
  - information gain (IG)
  - information gain ratio (IGR)
  - maximum entropy weight (MEW)

Feature Selection II

- Spatial measures:
  - TF-ICF (inverse city frequency)
  - geographical spread (geo) [Laere et al., 2014]
    1. divide the earth into $1^\circ \times 1^\circ$ cells
    2. calculate which cells word w occurs in
    3. merge neighbouring cells containing w until no more cells can be merged, and calculate the resulting number of cells containing w

$$GeoSpread(w) = \frac{\#\text{ of cells containing } w \text{ after merging}}{Max(w)}$$

where $Max(w) =$ max frequency of w in an unmerged cell.

Feature Selection III

- Ripley K function (Ripley) [Laere et al., 2014]: measures whether a given set of points is generated from a homogeneous Poisson distribution

$$K(\lambda) = A \times \frac{|\{p, q \in Q_w : \text{distance}(p, q) \leq \lambda\}|}{|Q_w|^2}$$

Source(s): Han et al. [2014]

Experimental Setup

- Datasets:
  - North America dataset (NA, Roller et al. [2012]) : 500K users, 38M tweets
  - World dataset (WORLD, Han et al. [2012b]): 1.4M users, 192M tweets
- Evaluation metrics:
  - Accuracy (Acc)
  - Accuracy within 161 KM (Acc@161), e.g., Frankfurt and Darmstadt
  - Country-level accuracy (Acc@C)
  - Median error distance
- Learner = multinomial naive Bayes (also experiments with logistic regression, incl. with L2 regulariser, but results worse)

Source(s): Han et al. [2014]

Multinomial Naive Bayes

- The basic formulation for multinomial NB is:

$$P(D|c_i) = \prod_{j=1}^{|\mathscr{V}|} \frac{P(t_j|c_i)^{N_{D,t_j}}}{N_{D,t_j}!}$$

where N_{D,t_j} is the frequency of the jth term in D, $\mathscr{V}$ is the set of all terms, and:

$$P(t|c_i) = \frac{1 + \sum_{k=1}^{|\mathscr{D}|} N_{k,t} P(c_i|D_k)}{|\mathscr{V}| + \sum_{j=1}^{|\mathscr{V}|} \sum_{k=1}^{|\mathscr{D}|} N_{k,t_j} P(c_i|D_k)}$$

- In practice, use addition of log-likelihoods rather than product of likelihoods

Source(s): McCallum and Nigam [1998]

Experimental Parameters

- Many variables and parameters to explore, including:
  - feature set: all tokens vs. LIWs
  - learner: multinomial naive Bayes (NB), Kullback-Leibler divergence (KL), maximum entropy (ME)
  - Location representation: City, k-D tree partitioned earth grid Roller et al. [2012]
  - language: English only, multilingual
  - data: geotagged data, geotagged and non-geotagged data, metadata

Source(s): Han et al. [2014]

Results using LIWs

Features	Acc	Acc@161	Acc@C	Median
Most Freq.	0.003	0.062	**0.947**	3089
Full	0.171	0.308	0.831	571
CHI	0.233	0.402	0.850	385
MaxCHI	0.238	0.412	0.848	356
LOGLIKE	0.191	0.343	0.836	489
IG	0.184	0.336	0.838	491
IGR	**0.260**	**0.450**	0.811	**260**
MEW	0.183	0.326	0.836	520
ICF	0.209	0.359	0.841	533
GEO	0.188	0.336	0.834	491
Ripley	0.236	0.432	0.849	306

Source(s): Han et al. [2014]

Models and Location Representation (NA)

Partition	Method	Acc	Acc@161	Acc@C	Median
k-D tree	KL	0.117	0.344	–	469
	KL+IGR	**0.161**	**0.437**	–	**273**
	NB	0.122	0.367	–	404
	NB+IGR	0.153	0.432	–	280
City	NB	0.171	0.308	0.831	571
	NB+IGR	**0.260**	**0.450**	0.811	**260**
	ME	0.129	0.232	0.756	878
	ME+IGR	0.229	0.406	0.842	369

* Acc is not comparable between different class representations

Source(s): Han et al. [2014]

Models and Location Representation (NA)

Partition	Method	Acc	Acc@161	Acc@C	Median
k-D tree	KL	0.117	0.344	–	469
	KL+IGR	**0.161**	**0.437**	–	**273**
	NB	0.122	0.367	–	404
	NB+IGR	0.153	0.432	–	280
City	NB	0.171	0.308	0.831	571
	NB+IGR	**0.260**	**0.450**	0.811	**260**
	ME	0.129	0.232	0.756	878
	ME+IGR	0.229	0.406	0.842	369

* Acc is not comparable between different class representations

Source(s): Han et al. [2014]

Models and Location Representation (NA)

Partition	Method	Acc	Acc@161	Acc@C	Median
k-D tree	KL	0.117	0.344	–	469
	KL+IGR	**0.161**	**0.437**	–	**273**
	NB	0.122	0.367	–	404
	NB+IGR	0.153	0.432	–	280
City	NB	0.171	0.308	0.831	571
	NB+IGR	**0.260**	**0.450**	0.811	**260**
	ME	0.129	0.232	0.756	878
	ME+IGR	0.229	0.406	0.842	369

* Acc is not comparable between different class representations

Source(s): Han et al. [2014]

Models and Location Representation (NA)

Partition	Method	Acc	Acc@161	Acc@C	Median
k-D tree	KL	0.117	0.344	–	469
	KL+IGR	**0.161**	**0.437**	–	**273**
	NB	0.122	0.367	–	404
	NB+IGR	0.153	0.432	–	280
City	NB	0.171	0.308	0.831	571
	NB+IGR	**0.260**	**0.450**	0.811	**260**
	ME	0.129	0.232	0.756	878
	ME+IGR	0.229	0.406	0.842	369

* Acc is not comparable between different class representations

Source(s): Han et al. [2014]

Models and Location Representation (NA)

Partition	Method	Acc	Acc@161	Acc@C	Median
k-D tree	KL	0.117	0.344	–	469
	KL+IGR	**0.161**	**0.437**	–	**273**
	NB	0.122	0.367	–	404
	NB+IGR	0.153	0.432	–	280
City	NB	0.171	0.308	0.831	571
	NB+IGR	**0.260**	**0.450**	0.811	**260**
	ME	0.129	0.232	0.756	878
	ME+IGR	0.229	0.406	0.842	369

* Acc is not comparable between different class representations

Source(s): Han et al. [2014]

Models and Location Representation (WORLD)

Dataset	Method	Acc	Acc@161	Acc@C	Median
City	NB	0.081	0.200	**0.807**	886
	NB+IGR	**0.126**	**0.262**	0.684	913
KD-tree	KL	0.116	0.283	-	564
	KL+IGR	0.121	0.286	-	602
	NB	0.119	0.289	-	**553**
	NB+IGR	**0.134**	**0.290**	-	577

Models and Location Representation (WORLD)

Dataset	Method	Acc	Acc@161	Acc@C	Median
City	NB	0.081	0.200	**0.807**	886
	NB+IGR	**0.126**	**0.262**	0.684	913
KD-tree	KL	0.116	0.283	-	564
	KL+IGR	0.121	0.286	-	602
	NB	0.119	0.289	-	**553**
	NB+IGR	**0.134**	**0.290**	-	577

Summary:

- Feature selection improves geolocation prediction accuracy
- Less impact of model and location representation choice than NA

Adding Non-geotagged Data

- In addition to the geotagged tweets from each user, we often have non-geotagged tweets, which we can potentially use to expand the training/test user representation

Train	Test	Acc	Acc@161	Acc@C	Median
G	G	0.126	0.262	0.684	913
G+NG	G	0.170	0.323	0.733	615
G	G+NG	0.187	0.366	0.835	398
G+NG	G+NG	0.280	0.492	0.878	170
G	G-small	0.121	0.258	0.675	960
G	NG-small	0.114	0.248	0.666	1057

Adding Non-geotagged Data

- In addition to the geotagged tweets from each user, we often have non-geotagged tweets, which we can potentially use to expand the training/test user representation

Train	Test	Acc	Acc@161	Acc@C	Median
G	G	0.126	0.262	0.684	913
G+NG	G	0.170	0.323	0.733	615
G	G+NG	0.187	0.366	0.835	398
G+NG	G+NG	0.280	0.492	0.878	170
G	G-small	0.121	0.258	0.675	960
G	NG-small	0.114	0.248	0.666	1057

- Incorporating NG improves the prediction accuracy
- There is minor difference between G-small and NG-small

Source(s): Han et al. [2014]

Exploration of Language Influence

- All results to date on English data; some languages highly predictive of location (e.g. Japanese, Finnish)
- Investigate interaction between language and geolocation accuracy:
  - Partition: city
  - Learner: multinominal naive Bayes
  - Training: IGR on geotagged multilingual data

Method	Acc	Acc@161	Acc@C	Median
Per-language majority classes	0.107	0.189	0.693	2805
Unified multilingual model	0.196	0.343	0.772	466
Monolingual partitioned model	**0.255**	**0.425**	**0.802**	**302**

Table : WORLD in multilingual settings.

Exploration of Language Influence

- All results to date on English data; some languages highly predictive of location (e.g. Japanese, Finnish)
- Investigate interaction between language and geolocation accuracy:
  - Partition: city
  - Learner: multinominal naive Bayes
  - Training: IGR on geotagged multilingual data

Method	Acc	Acc@161	Acc@C	Median
Per-language majority classes	0.107	0.189	0.693	2805
Unified multilingual model	0.196	0.343	0.772	466
Monolingual partitioned model	**0.255**	**0.425**	**0.802**	**302**

Table : WORLD in multilingual settings.

Language is a good indicator of location (EN hard!)

Source(s): Han et al. [2014]

User Metadata in Tweets

- Examples of user-declared location in public profile:
  - *Calgary, Alberta*
  - *Iowa, USA*
  - *heat of Arizona*
  - *north east side of indy*
  - *-iN A Veryy Dope Place (:*
  - *hugging my big sister*
- Examples of user-declared real names in public profile:
  - *Michael Jordan*
  - *Yuji Matsumoto*
  - *Hinrich Schutze*

User Metadata in Tweets

- Examples of user-declared location in public profile:
  - *Calgary, Alberta*
  - *Iowa, USA*
  - *heat of Arizona*
  - *north east side of indy*
  - *-iN A Veryy Dope Place (:*
  - *hugging my big sister*
- Examples of user-declared real names in public profile:
  - *Michael Jordan*
  - *Yuji Matsumoto*
  - *Hinrich Schutze*

Train NB classifier for each of user-declared location, timezone, self-description, registered real name

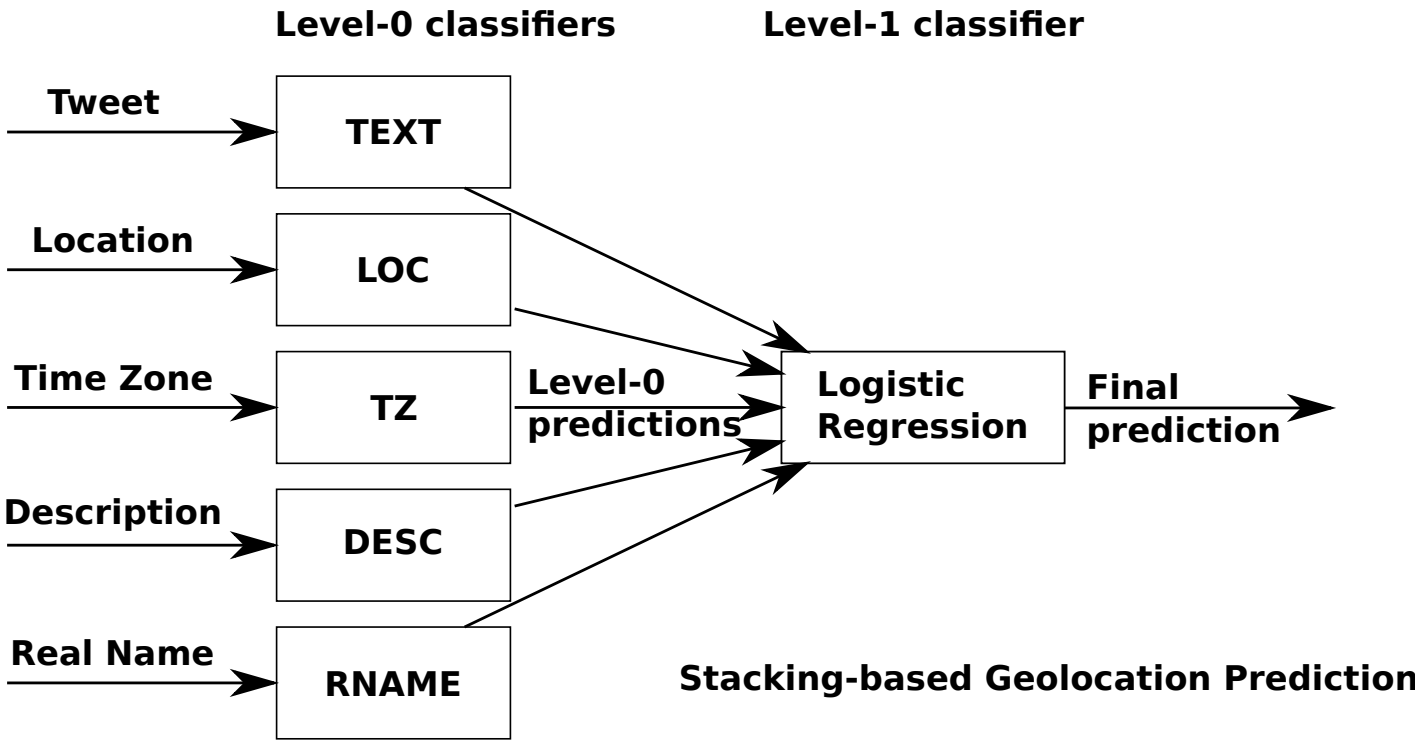
Source(s): Han et al. [2014]

Exploration of User Metadata (WORLD)

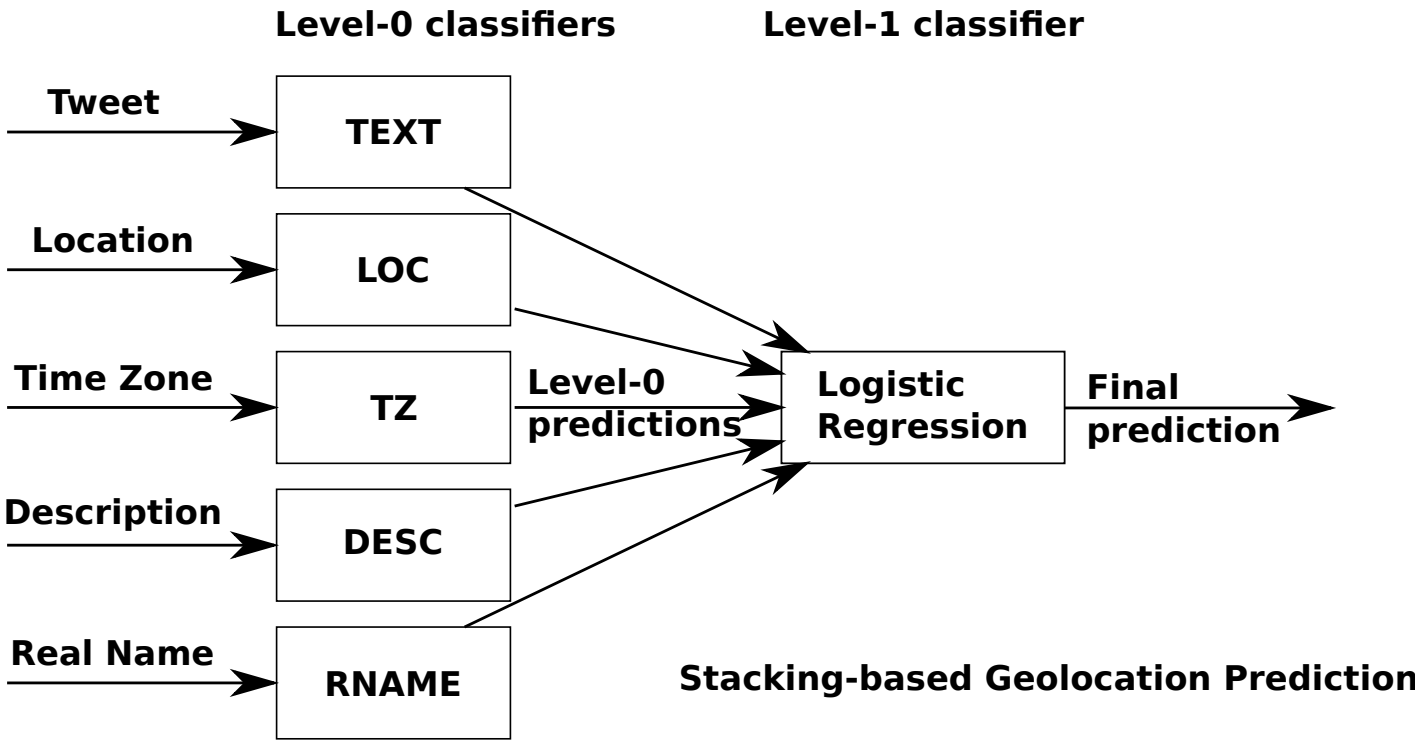
Classifier	Acc	Acc@161	Acc@C	Median
TEXT	0.280	0.492	0.878	170
LOC	0.405	0.525	0.834	92
TZ	0.064	0.171	0.565	1330
DESC	0.048	0.117	0.526	2907
RNAME	0.045	0.109	0.550	2611

Source(s): Han et al. [2014]

Stacking Metadata Classifiers

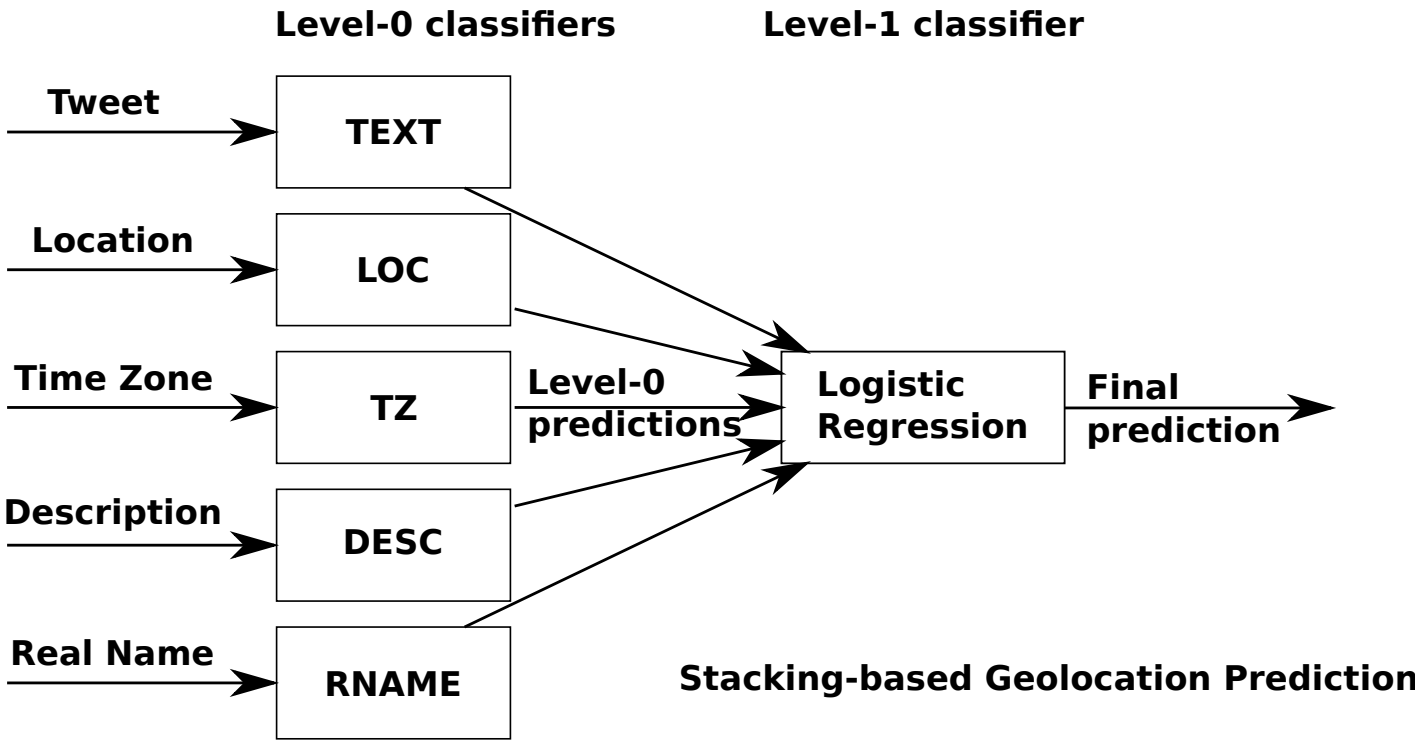


Stacking Metadata Classifiers



	Features	Acc	Acc@161	Acc@C	Median
0.	TEXT	0.280	0.492	0.878	170
1.	0. + LOC	0.483	0.653	0.903	14
2.	1. + TZ	0.490	0.665	0.917	9
3.	2. + DESC	0.490	0.666	0.919	9
4.	3. + RNAME	0.491	0.667	0.919	9

Stacking Metadata Classifiers



	Features	Acc	Acc@161	Acc@C	Median
0.	TEXT	0.280	0.492	0.878	170
1.	0. + LOC	0.483	0.653	0.903	14
2.	1. + TZ	0.490	0.665	0.917	9
3.	2. + DESC	0.490	0.666	0.919	9
4.	3. + RNAME	0.491	0.667	0.919	9

Temporal Influence

- Can a model trained on "old" data generalise to "new" data?

WORLD: 10K time-homogeneous users

Features	Acc	Acc@161	Acc@C	Median
1. TEXT	0.280	0.492	0.878	170
2. LOC	0.405	0.525	0.834	92
3. TZ	0.064	0.171	0.565	1330
1. + 2. + 3.	0.490	0.665	0.917	9

LIVE: 32K time-heterogeneous users

Features	Acc	Acc@161	Acc@C	Median
1. TEXT	0.268	0.510	0.901	151
2. LOC	0.326	0.465	0.813	306
3. TZ	0.065	0.160	0.525	1529
1. + 2. + 3.	0.406	0.614	0.901	40

Source(s): Han et al. [2014]

Temporal Influence

- Can a model trained on "old" data generalise to "new" data?

WORLD: 10K time-homogeneous users

Features	Acc	Acc@161	Acc@C	Median
1. TEXT	0.280	0.492	0.878	170
2. LOC	0.405	0.525	0.834	92
3. TZ	0.064	0.171	0.565	1330
1. + 2. + 3.	0.490	0.665	0.917	9

LIVE: 32K time-heterogeneous users

Features	Acc	Acc@161	Acc@C	Median
1. TEXT	0.268	0.510	0.901	151
2. LOC	0.326	0.465	0.813	306
3. TZ	0.065	0.160	0.525	1529
1. + 2. + 3.	0.406	0.614	0.901	40

Source(s): Han et al. [2014]

Prediction Confidence I

More geolocatable user:

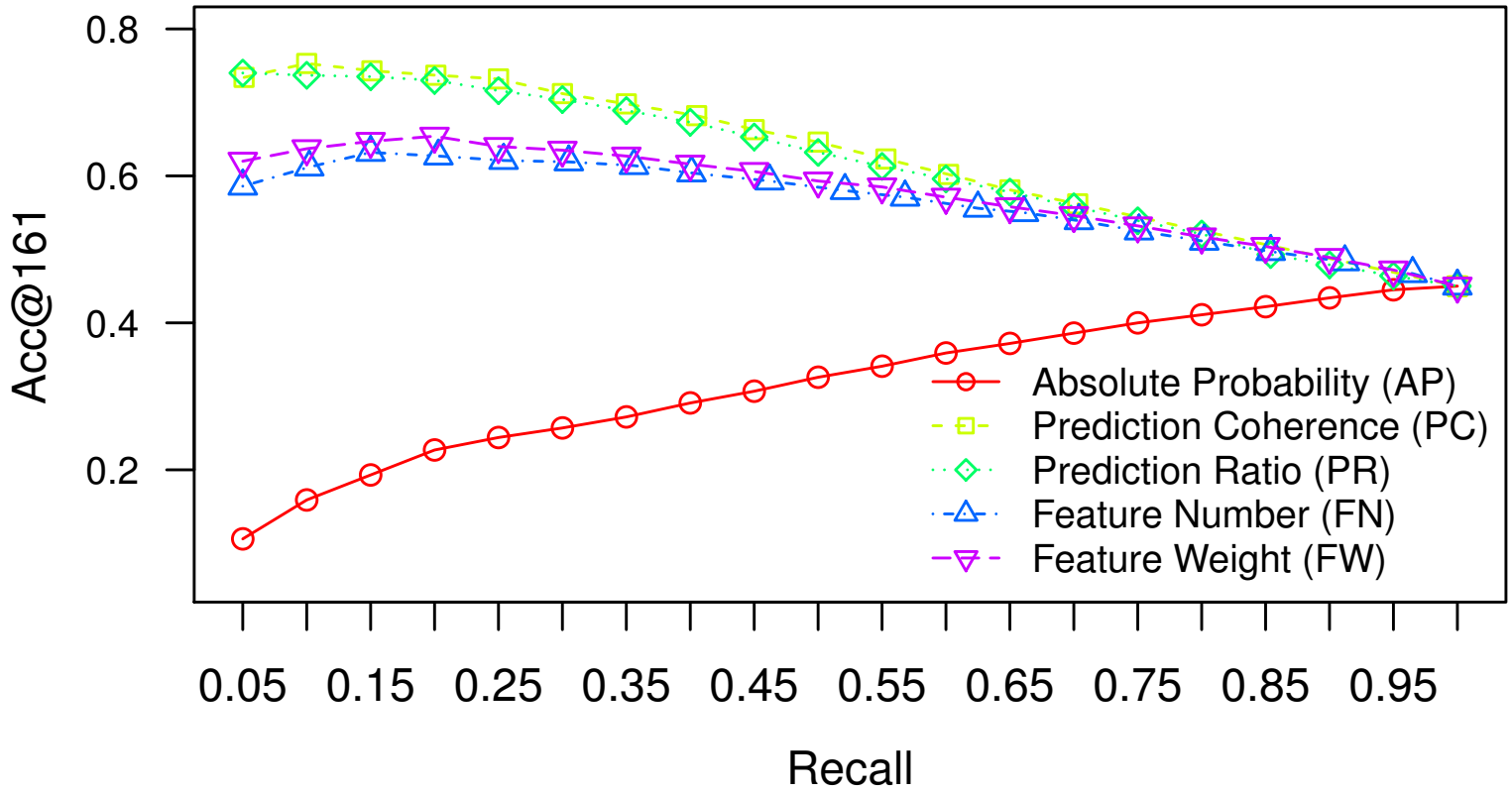
- Porting my mobile to Telstra is a brilliant idea, #vodafail
- @USER1 @USER2 @USER3 actually Kevin Rudd also has an active weibo account.
- @USER good memory, I can hardly remember the day I came to Melbourne.

Less geolocatable user:

- happy birthday to me
- i just finished my hw, oooh, too much
- Yes! all things are diffcult before they re easy

Prediction Confidence II

- Rank users by confidence: probability (AP), probability ratio of 1st and 2nd prediction (PR), geo-proximity in top-10 predictions (PC), accumulated counts (FN) and weights (FW) of optimised features:



Source(s): Han et al. [2014]

The Road Ahead

- Network features highly effective in user geolocation (moreso than text features); much work to be done in combining the two
- Message-level geolocation still very much an unsolved task
- How to keep the model temporally-relevant?
- Interaction between lexical normalisation and geolocation

Summary

- User geolocation: supervised text-based multi-classification problem
- Location Indicative Words improve model effectiveness and efficiency
- Model choice and location partitions are less crucial than feature selection
- Adding non-geotagged data, identifying languages and incorporating user metadata all improve the prediction accuracy

Talk Outline

Twitter POS Tagging

- How is POS tagging for social media data (focusing on Twitter) different to POS tagging for any other text source?
  - deterministically taggable tokens (URLs, emoticons)
  - higher proportion of OOV words → lexical normalisation, beef-up novel word handling rules, add word cluster information
  - lower reliability of casing → add more gazetteers
  - lots of untagged, little tagged data → incorporate semi-supervised retraining (e.g. bootstrapping)
  - some POS tag distinctions hard to make in social media → tweak POS tagset to remove certain distinctions (and add others)

Source(s): Gimpel et al. [2011a], Derczynski et al. [2013], Owoputi et al. [2013]

References I

AiTi Aw, Min Zhang, Juan Xiao, and Jian Su. A phrase-based statistical model for SMS text normalization. In *Proceedings of COLING/ACL 2006*, pages 33–40, Sydney, Australia, 2006.

Eric Brill and Robert C. Moore. An improved error model for noisy channel spelling correction. In *ACL '00: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, pages 286–293, Hong Kong, 2000.

Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. Investigation and modeling of the structure of texting language. *International Journal on Document Analysis and Recognition*, 10:157–174, 2007.

Paul Cook and Suzanne Stevenson. An unsupervised model for text message normalization. In *CALC '09: Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*, pages 71–78, Boulder, USA, 2009.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of RANLP 2013 (Recent Advances in Natural Language Processing)*, Hissar, Bulgaria, 2013.

Michelle Drouin and Claire Davis. R u txting? is the use of text speak hurting your literacy? *Journal of Literacy Research*, 41(1):46–67, 2009.

Jacob Eisenstein. What to do about bad language on the internet. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 359–369, Atlanta, USA, 2013. URL `http://www.aclweb.org/anthology/N13-1037`.

References II

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP 2010)*, pages 1277–1287, Cambridge, USA, 2010. URL `http://www.aclweb.org/anthology/D10-1124`.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 42–47, Portland, USA, 2011a. URL `http://www.aclweb.org/anthology/P11-2008`.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 42–47, Portland, USA, 2011b.

Stephan Gouws, Dirk Hovy, and Donald Metzler. Unsupervised mining of lexical variants from noisy text. In *Proceedings of the First workshop on Unsupervised Learning in NLP*, pages 82–90, Edinburgh, UK, 2011a.

Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 20–29, Portland, USA, 2011b. URL `http://www.aclweb.org/anthology/W11-0704`.

References III

Bo Han and Timothy Baldwin. Lexical normalisation of short text messages: Makn sens a #twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 368–378, Portland, USA, 2011.

Bo Han, Paul Cook, and Timothy Baldwin. Automatically constructing a normalisation dictionary for microblogs. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2012 (EMNLP-CoNLL 2012)*, pages 421–432, Jeju, Korea, 2012a.

Bo Han, Paul Cook, and Timothy Baldwin. Geolocation prediction in social media data by finding location indicative words. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1045–1062, Mumbai, India, 2012b.

Bo Han, Paul Cook, and Timothy Baldwin. Text-based Twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 49:451–500, 2014.

Olivier Van Laere, Jonathan Quinn, Steven Schockaert, and Bart Dhoedt. Spatially-aware term selection for geotagging. *IEEE Transactions on Knowledge and Data Engineering*, 26(1):221–234, 2014.

Wang Ling, Chris Dyer, Alan W Black, and Isabel Trancoso. Paraphrasing 4 microblog normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 73–84, Seattle, USA, 2013. URL `http://www.aclweb.org/anthology/D13-1008`.

References IV

Fei Liu, Fuliang Weng, Bingqing Wang, and Yang Liu. Insertion, deletion, or substitution? normalizing text messages without pre-categorization nor supervision. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 71–76, Portland, USA, 2011. URL `http://www.aclweb.org/anthology/P11-2013`.

Andrew McCallum and Kamal Nigam. A comparison of event models for Naive Bayes text classification. In *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, pages Available as Technical Report WS–98–05, AAAI Press., Madison, USA, 1998.

Olutobi Owoputi, Brendan OConnor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2013)*, pages 380–390, Atlanta, USA, 2013.

Lawrence Philips. The Double Metaphone Search Algorithm. *C/C++ Users Journal*, 18: 38–43, 2000. ISSN 1075-2838.

Reid Priedhorsky, Aron Culotta, and Sara Y. Del Valle. Inferring the origin locations of tweets with quantitative confidence. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW 2014)*, pages 1523–1536, Baltimore, USA, 2014.

References V

Stephen Roller, Michael Speriosu, Sarat Rallapalli, Benjamin Wing, and Jason Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning 2012 (EMNLP-CoNLL 2012)*, pages 1500–1510, Jeju Island, Korea, 2012. URL `http://www.aclweb.org/anthology/D12-1137`.

Yi Yang and Jacob Eisenstein. A log-linear model for unsupervised text normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*, pages 61–72, Seattle, USA, 2013. URL `http://www.aclweb.org/anthology/D13-1007`.

Congle Zhang, Tyler Baldwin, Howard Ho, Benny Kimelfeld, and Yunyao Li. Adaptive parser-centric text normalization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, pages 1159–1168, Sofia, Bulgaria, 2013.