Contextual Bearing on Linguistic Variation in Social Media

Stephan Gouws
USC Information Sciences Institute

Joint work with:
Donald Metzler, Congxing Cai and Eduard Hovy

Introduction

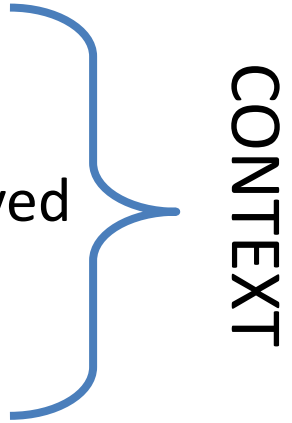
- Ubiquitous nature of Internet-enabled devices leads to large volumes of user-generated content
- Language usage in these messages vary wildly:

  @amandadiva: “no matter what ya'll say or do gotta get this paperrrrrr... night ya'll.tmrw we rise and grind (as usual)”

  “@Chrisssyyyyyy: #shoutout nd thaank yu 2 all dose who sed Happy birthday, God Bless nd ill be sure 2 say hbd back wen its yurs :) my sweet wittle thangs”

Lexical Variation, leksikal Variashn, lexcl vartn..

- The result of several factors
  - social factors
    - age, location, how they want to be perceived
  - usability & interface issues
    - unwieldy keypads, length constraints



- We look at a user's:
  - **location** and
  - **device**

Motivation

- Off-the-shelf NLP tools break on ill-formed text
  - Parsing, NER, ...
- Real-time nature affords unique view on users and their perspectives
- Strong motivation to analyse this type of text
- Yet word variations might skew analyses
- Nonstandard orthography is **intentional**

Types of lexical variation

@amandadiva: “no matter what ya'll say or do gotta get this paperrrrrr... night ya'll.tmrw we rise and grind (as usual)”

“@Chrisssyyyyyy: #shoutout nd thaank yu 2 all dose who sed Happy birthday, God Bless nd ill be sure 2 say hbd back wen its yurs :) my sweet wittle thangs”

Types of lexical compression schemes

Transformation Type	Example
Single-character	“see” -> “c”
Suffix-style compression	“about”-> “bout”
Drop vowels	“and” -> “nd”
Prefix-style compression	“tomorrow” -> “tom”
You-to-u	“you” -> “u”
Character clipping	“running” -> “runnin”
Repeat letter	“thank” -> “thaank”
Contractions	“you all” -> “y’all”
Th-to-d	“those” -> “dose”

Goal

Analyse influence of these contexts on lexical structure/orthography of messages

Our Approach

Two types of analyses:

1. Naive OOV analysis
2. More sophisticated analysis using text normalisation

Our Approach

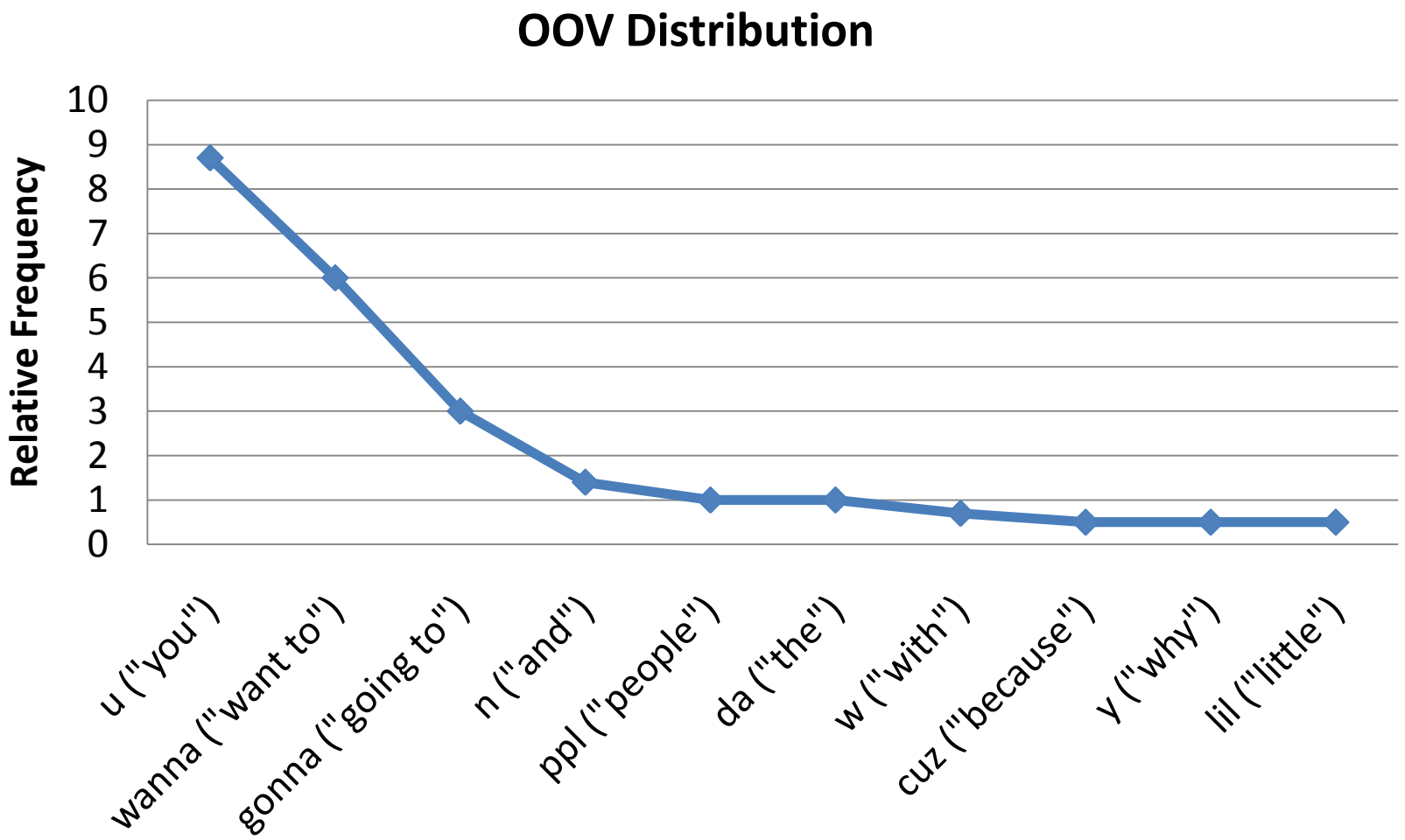
Two types of analyses:

1. **Naive OOV analysis**
2. More sophisticated analysis using text normalisation

OOV Analysis

- Conducted over 1M tweets
- Lexicon extracted from LA Times (>500K)
- OOVs largely (>50%) consist of usernames, hash tags, “RT”, “LOL”, “haha”, “lmao” + numerous expletives

OOV Analysis: Most frequent OOVs



OOV Analysis

Timezone	% In-Vocabulary
Australia	86%
UK	85%
US (Atlantic)	84%
Hong Kong	83%
US (Pacific)	81%
Hawaii	81%
Overall	81%

Australia and UK write more “proper” tweets than the rest of the world

Our Approach

Two types of analyses:

1. Naive OOV analysis
2. **More sophisticated analysis using text normalisation**

Why normalise?

- "@Chrisssyyyyyy: #shoutout nd thaank yu 2 all dose who sed Happy birthday
- OOV words: nd, thaank, yu, 2, dose, sed...
  - nd –> and [drop vowels, suffix]
  - thaank –> thank [repeat letter]
  - yu –> you [drop vowels]
  - ...
- **Question**: How is user's writing style influenced by 1) location, and 2) device?

Noisy Text Normalisation

- “c u tmrw” -> “see you tomorrow”
- Lot of attention recently:
  - Spelling correction [Choudhury 2007, Cook+Stevenson 2009]
  - MT [Aw et al. 2006, Kaufman+Kalita 2010],
  - ASR [Kobus 2008]
  - Hybrid [Han et al. 2011]

Noisy Text Normalisation
Our hybrid approach

1) OOV detection
2) Candidate enumeration
3) Decoding

Noisy Text Normalisation

Our hybrid approach

1) OOV detection

2) Candidate enumeration

3) Decoding

@Chrisssyyyyyy #shoutout nd thaank yu 2 all dose who sed Happy birthday

Noisy Text Normalisation

Our hybrid approach

1) OOV detection -> Heuristics + Lexicon lookup

2) Candidate enumeration

3) Decoding

@Chrisssyyyyyy #shoutout nd thaank yu 2 all dose who sed Happy birthday

Noisy Text Normalisation

Our hybrid approach

1) OOV detection -> Heuristics + Lexicon lookup

2) Candidate enumeration

3) Decoding

@Chrisssyyyyyy #shoutout nd thaank yu 2 all ~~dose~~ who sed Happy birthday

Problem?

When transformations/errors transform a word to an IV word, e.g. “those”->“dose”

Noisy Text Normalisation

Our hybrid approach

1) OOV detection

2) Candidate enumeration

3) Decoding

@Chrisssyyyyyy #shoutout nd thaank yu 2 all ~~dose~~ who sed Happy birthday

Noisy Text Normalisation

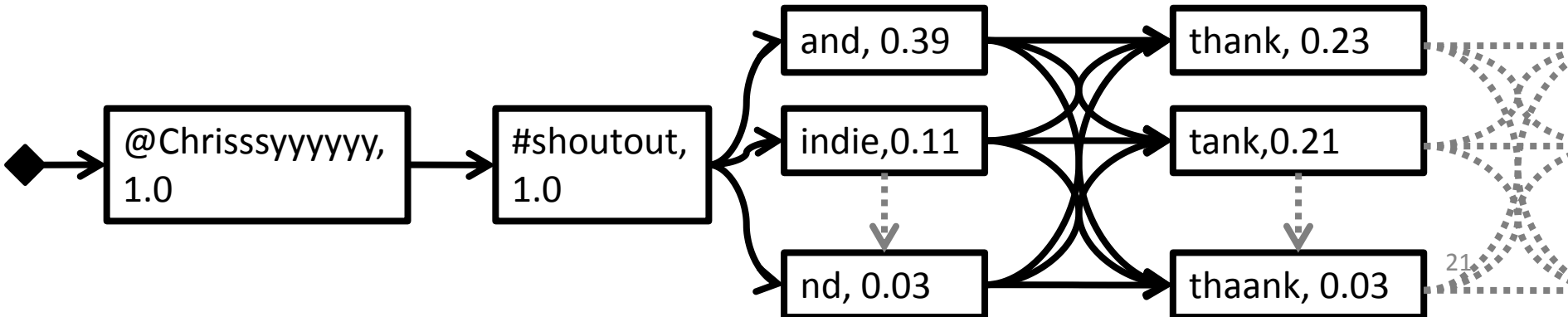
Our hybrid approach

1) OOV detection

2) Candidate enumeration -> Confusion network

3) Decoding

@Chrisssyyyyyy #shoutout nd thaank yu 2 all ~~dose~~ who sed Happy birthday



Noisy Text Normalisation

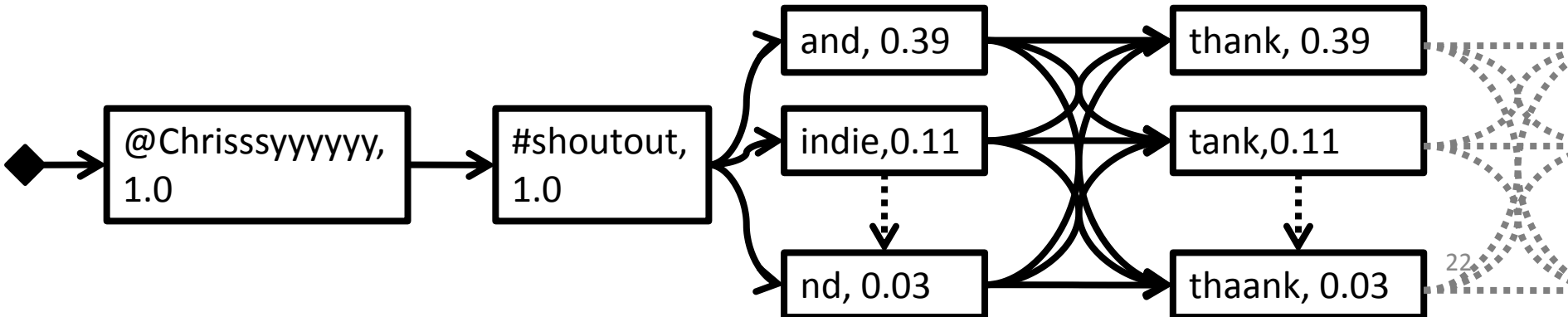
Our hybrid approach

1) OOV detection

2) Candidate enumeration

3) Decoding -> maximum posterior path

@Chrisssyyyyyy #shoutout nd thaank yu 2 all ~~dose~~ who sed Happy birthday



Noisy Text Normalisation

Our hybrid approach

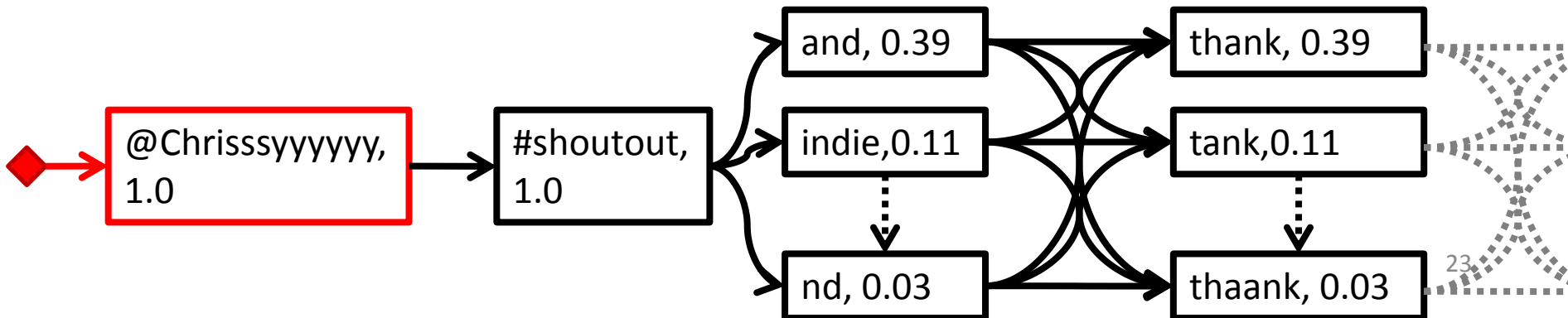
1) OOV detection

2) Candidate enumeration

3) Decoding -> posterior Viterbi path

@Chrisssyyyyyy

@Chrisssyyyyyy #shoutout nd thaank yu 2 all ~~dose~~ who sed Happy birthday



Noisy Text Normalisation

Our hybrid approach

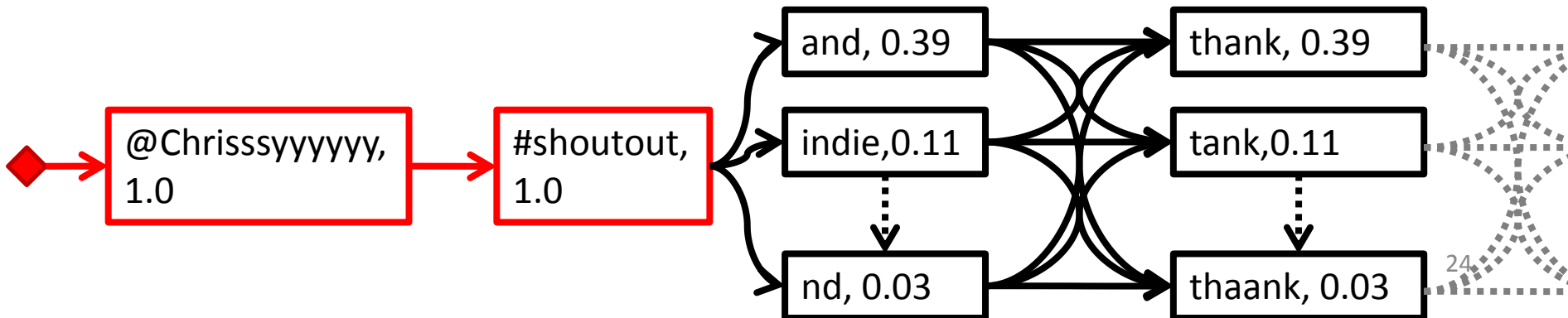
1) OOV detection

2) Candidate enumeration

3) Decoding -> posterior Viterbi path

@Chrisssyyyyyy #shoutout

@Chrisssyyyyyy #shoutout nd thaank yu 2 all ~~dose~~ who sed Happy birthday



Noisy Text Normalisation

Our hybrid approach

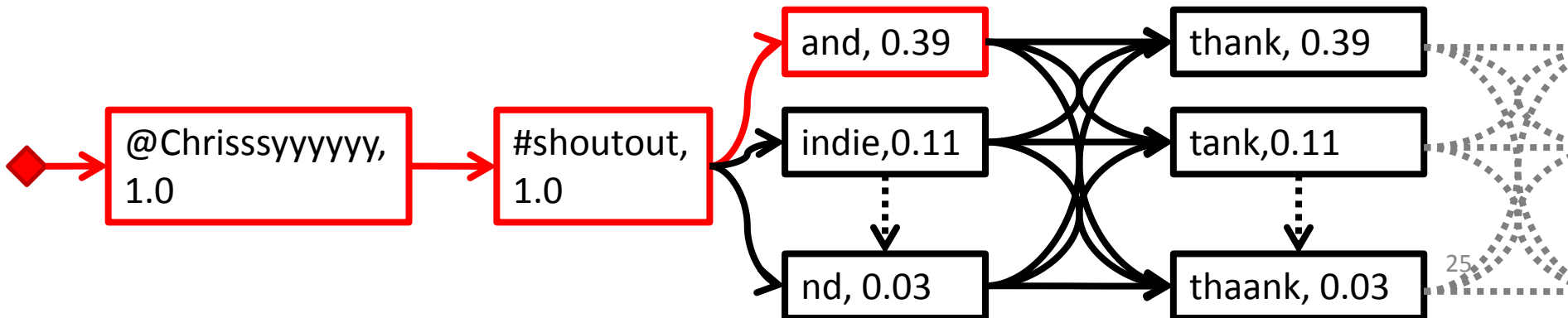
1) OOV detection

2) Candidate enumeration

3) Decoding -> posterior Viterbi path

@Chrisssyyyyyy #shoutout and

@Chrisssyyyyyy #shoutout nd thaank yu 2 all ~~dose~~ who sed Happy birthday



Noisy Text Normalisation

Our hybrid approach

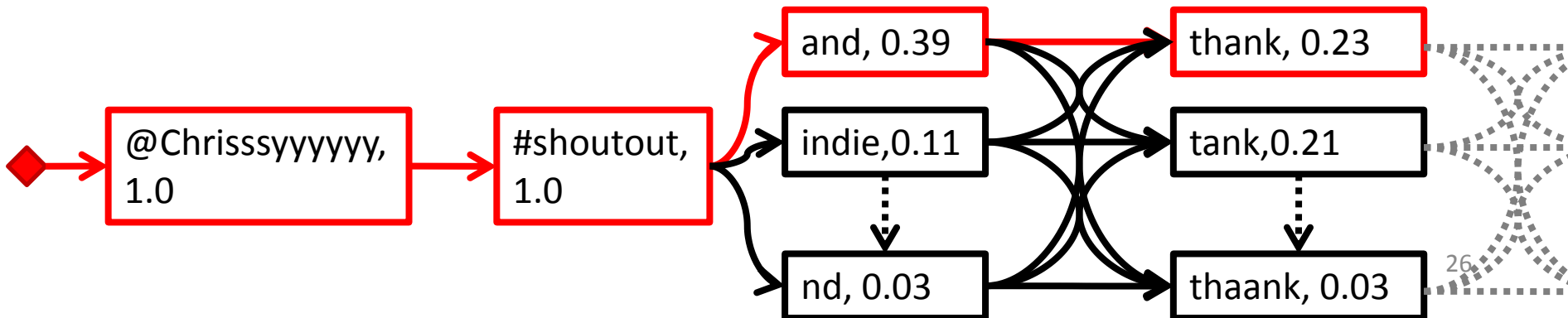
1) OOV detection

2) Candidate enumeration

3) Decoding -> posterior Viterbi path

@Chrisssyyyyyy #shoutout and thank ...

@Chrisssyyyyyy #shoutout nd thaank yu 2 all ~~dose~~ who sed Happy birthday



Detailed Analysis

- **Question**: How is user's writing style influenced by:
  - geo-location (Hawaii, Pacific US, London, and Australia)
  - platform (iPhone, BlackBerry, Web (PC), UberTwitter for Android)

Detailed Analysis

1. Normalise noisy text
2. Record types of transformations
3. Analyse the types of transformations that occur in the different contexts

Results by Device

Device	...exhibit much more...	...than device
	Contractions “you will” -> “you’ll”	
	Drop vowels “about” -> “abt”	 + 
 + 	Suffix-style compressions “tomorrow” -> “morrow”	

Results by Location

Location	...exhibit much more...	...than location
	Contractions, and Suffix-style compressions	
	Drop last letter, Drop vowels, Prefix-style compressions, Repeat letters	

Discussion

- A few types represent > 50% of OOV tokens
  - Few simple deterministic rules can clean up large amount of errors
- Empirically verified types of transformations vary across location and device
  - NLP applications for social media -> context-aware

Conclusion

- Analysed writing styles of different users found in microtexts
- Showed how writing styles change across users' locations and platforms
- Impacts this has on HLT applications in Social Media

Thanks!

- Questions?

Extras

Lexical Transformations

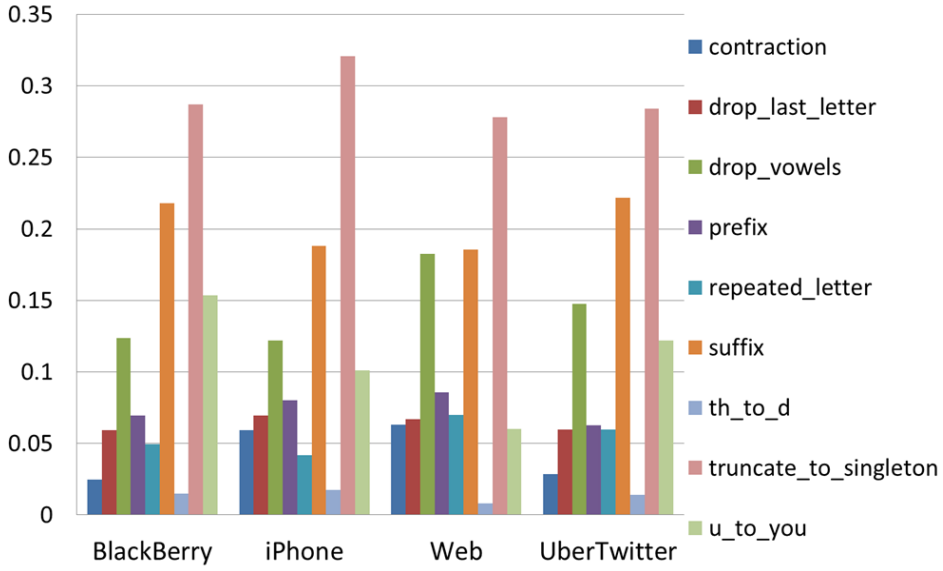
Transformation Type	Example	Frequency
Single-character	“see” -> “c”	29.1%
Suffix-style compression	“about”-> “bout” “because” -> “cause”	18.8%
Drop vowels	“and” -> “nd”	16.4%
Prefix-style compression	“tomorrow” -> “tom”	9.0%
You-to-u	“you” -> “u”	8.3%
Character clipping	“running” -> “runnin”	7.0%
Repeat letter	“thank” -> “thaank”	5.5%
Contractions	“you all” -> “y’all”	5.0%
Th-to-d	“those” -> “dose”	1.0%

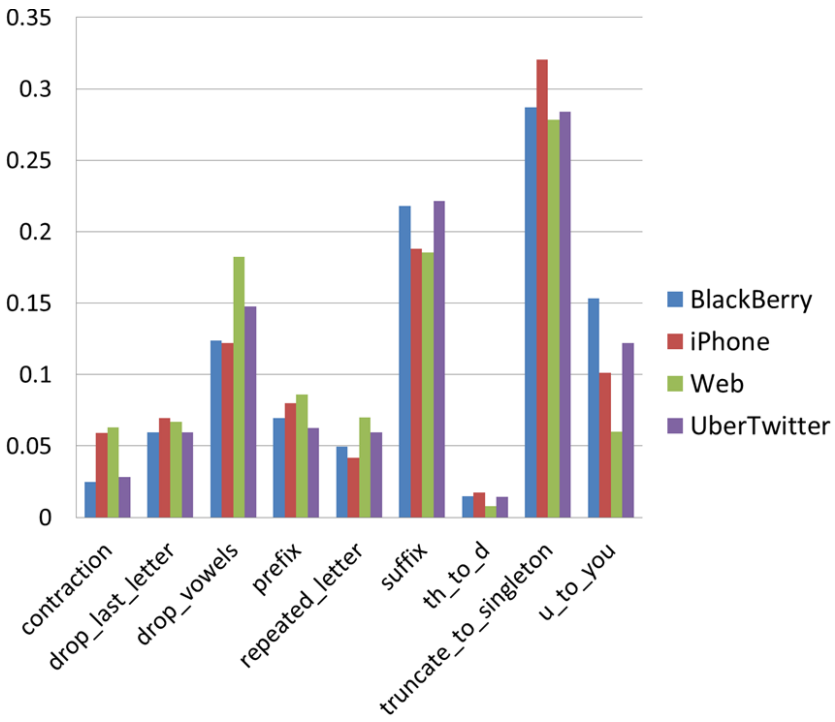
OOV Analysis by Platform

Client	% In-Vocabulary
Facebook	88%
Twitter for iPhone	84%
Twitter for Blackberry	83%
Web	82%
UberTwitter	78%
Snaptu	73%
Overall	81%

- Why Facebook so “clean”?
- iPhone/BB > Web

Contextual Analysis by Platform





Contextual Analysis by Location