Code-Mixing: A Challenge for Language Identification in the Language of Social Media

Utsab Barman, Amitava Das[♠], Joachim Wagner & Jennifer Foster

Dublin City University,
Dublin, Ireland.

[♠] University of North Texas,
Denton, USA.

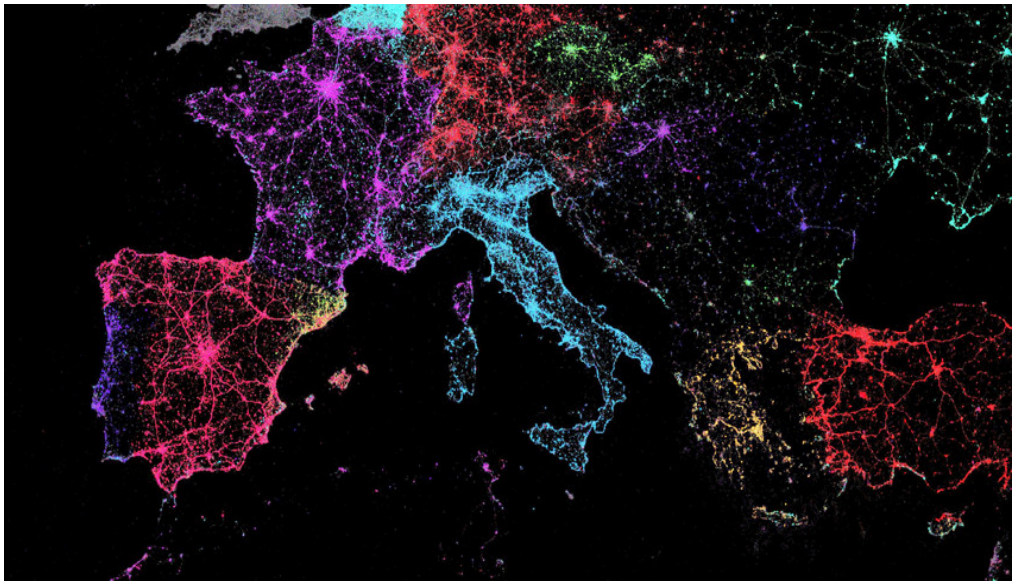
DATE – 25.10.2014

WWW.CNGL.IE

Language Identification in Social Media is a Challenging Task

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Twitter Language Map



http://www.fastcodesign.com/1665366/infographic-of-the-day-the-many-languages-of-twitter

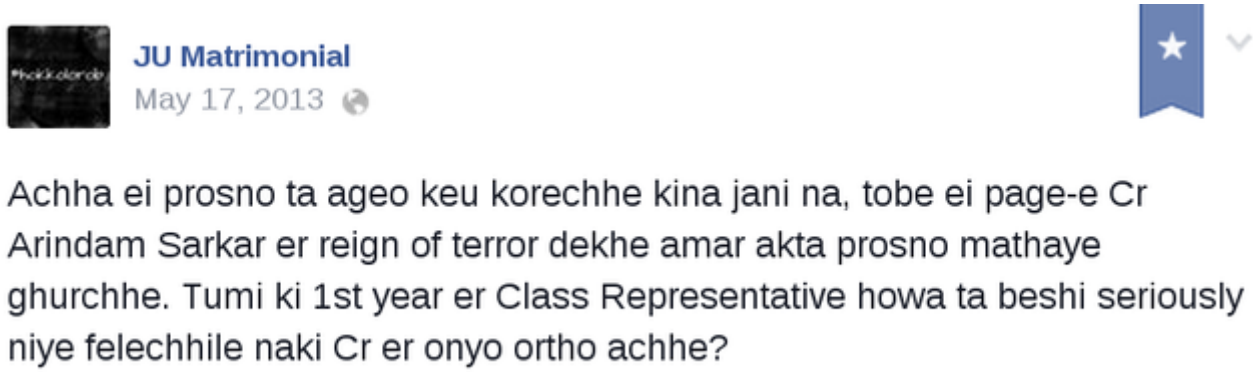
- Plenty of languages
  - Only half of them are in English
- Informal writing
  - Great -> gr8
- Code-mixing

Code-Mixing

CENTRE FOR GLOBAL INTELLIGENT CONTENT

- Mixing multiple languages
  - Inter-sentential
  - Intra-sentential
  - Word-level
- Phonetic typing
  - Writing in Roman script instead of native language script
  - Ad-hoc Romanisation

Example : Phonetically Typed Code-Mixed Content



Achha ei prosno ta ageo keu korechhe kina jani na, tobe ei page-e Cr Arindam Sarkar er reign of terror dekhe amar akta prosno mathaye ghurchhe. Tumi ki 1st year er Class Representative howa ta beshi seriously niye felechhile naki Cr er onyo ortho achhe?

- Bengali
- English

Goal of our Work

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Word-Level Language Identification with Phonetically Typed Code-Mixed Content

Corpus

- **English-Hindi-Bengali**
- phonetically typed code-mixed content
- Facebook post and comments
- Indian student community

Reasons:

- Code-mixing is frequent among speakers who are multilingual and younger in age.
  - India is a country with 30 spoken languages, among which 22 are official.
  - 65% of Indian population is 35 or under. **

Currently our corpus contains 2335 posts and 9813 comments.

** http://www.theguardian.com/commentisfree/2014/apr/08/india-leaders-young-people-change-2014-elections

Annotation (1)

- **Annotation Type: Human Annotation**

- **Number of Annotators: 4**
  - 3 students from Computer Science background from same university
  - 1 author of this paper

- **Target:**
  - Capture
    - inter-sentential code-mixing
    - intra-sentential code-mixing
    - word-level code-mixing

Annotation (2)

- **Tags: <T attribute = “L”> </T>**
  - **T**: Type of cde-mixing
    - sentence (sent)
    - fragment (frag)
    - inclusion (incl)
    - word level code-mixing (wlcm)
  - **L**: Language(s) of code-mixing
    - English (en)
    - Hindi (hi)
    - Bengali (bn)
    - Mixed (mixd)
    - Universals (univ)
    - Undefined (undef)

Annotation (3)

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Sentence

<sent lang =”language”> .... </sent>

- Identifies sentence boundary
- Identifies inter-sentential code-mixing

Annotation (4)

- **English Sentence:** **what a.....6 hrs long...but really nice tennis....**

  <sent lang="**en**">

  what a.....6 hrs long...but really nice tennis....

  </sent>

- **Bengali Sentence:** **shubho nabo borsho.. :)**

  <sent lang="**bn**">

  shubho nabo borsho.. :)

  </sent>

- **Hindi Sentence:** **karwa sachh ..... :(**

  <sent lang="**hi**">

  karwa sachh ..... :(

  </sent>

Annotation (5)

- **Univ-Sentence: hahahahahahah....!!!!!**

  <sent lang="univ">

  hahahahahahah....!!!!!

  </sent>

- **Mixed-Sentence: oye hoye ..... angreji me kahte hai ke I love u.. !!!**

  <sent lang="mixd">

  <frag lang="**hi**">

  oye hoye ..... angreji me kahte hai ke

  </frag>

  <frag lang="**en**">

  I love u.. !!!

  </frag>

  </sent>

Annotation (6)

- **Fragment**

  <frag lang =”language”> ... </frag>

  - Identifies groups of grammatically related words in a sentence
  - Identifies intra-sentential code-mixing

Annotation (7)

Mixed-Sentence: oye hoye ..... angreji me kahte hai ke I love u.. !!!

```
<sent lang="mixd">
    <frag lang="hi">
        oye hoye ..... angreji me kahte hai ke
    </frag>
    <frag lang="en">
        I love u.. !!!
    </frag>
</sent>
```

Annotation (8)

Inclusion

<incl lang=”language”> ... </incl>

- Identifies foreign word or phrase
- Within sentence or fragment
- Assimilated in native language
- Identifies intra-sentential code-mixing

Annotation (9)

Sentence with inclusion: **Na re seriously ami khub kharap achi.**

```
<sent lang=”bn”>
    Na re
    <incl lang=”en”>
        seriously
    </incl>
    ami khub kharap achi.
</sent>
```

Annotation (10)

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Word-Level Code-Mixing

<wlcm type="languages"> ... </wlcm>

- Capture intra-word code-mixing
- Smallest unit of code-mixing

Annotation (11)

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Word-level code mixing (EN-BN) : chapless

where

- **Root word: chap (Bengali)**
- **Appended Suffix: less (English)**

```
<wlcm type="bn-and-en">
    chapless
</wlcm>
```

Token-Level Statistics

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Language	Count
EN	66,298
BN	79,899
HI	3,440
WLCM	633
UNIV	39,291
UNDEF	61

5,233 tokens are identified as NE and 715 tokens are identified as Acronym (e.g. JU).

Total: 195,570

Tag-Level Statistics

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Tags	EN	BN	HI	Mixd	Univ	Undef
sent	5,370	5,523	354	204	746	15
frag	288	213	40	-	6	0
incl	7,377	262	94	-	1,032	1
wlcm	477					

Ambiguous Words

Labels	Count	Percentage
EN	9,109	34.40
BN	14,345	54.18
HI	1,039	3.92
EN or BN	1,479	5.58
EN or HI	61	0.23
BN or HI	277	1.04
EN or BN or HI	165	0.62

- Some types are annotated in multiple languages,
  e.g 'to', 'clg', 'baba'
  - Common vocabulary between languages
  - Effect of phonetic typing

IAA (1)

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Token-Level Kappa = 0.884

[Calculated on randomly selected 100 comments between 2 annotators]

IAA (2)

- Tag-level Kappa = 0.6683

Tag	Kappa
sent	0.6825
frag	0.5171
incl	0.5507
wlcm	0.6223
ne	0.6172
acro	0.6144
All tags	0.6683

Annotation

<sent lang="bn">ki
<incl lang="en"> cntrl </incl>
korte parli na
</sent>

Word-level representation

B-SENT-bn	ki
B-INCL-en/I-SENT-bn	cntrl
I-SENT-bn	krte
I-SENT-bn	parli
I-SENT-bn	na

[Calculated on randomly selected 100 comments between 2 annotators]

Experiments (1)

CENTRE FOR GLOBAL INTELLIGENT CONTENT

- Approaches
  - Dictionary-based
  - SVM without contextual information
  - SVM and CRF with contextual information

- 5-fold cross-validation
- 4-way classification (en, bn, hi and univ)

Experiments (2)

- To avoid unrealistic context,
  - NEs and WLCMs are included for context features
  - With label 'other' in training (5-way system)
- Two special cases:
  - Gold NEs and WLCMs do not count for evaluation
  - Back-off to 4-way system (en, bn, hi and univ) when 'other' is predicted

Dictionary Approach (1)

- Full-form dictionaries extracted from
  - British National Corpus
  - SEMEVAL 2013 Twitter data
  - Lexical normalisation list (Han and Baldwin, 2011)
  - Training data
- No transliterated Bengali or Hindi dictionary available

Dictionary Approach (2)

CENTRE FOR GLOBAL INTELLIGENT CONTENT

- Language prediction by presence in dictionaries
- Use normalised word frequencies
- For OOVs or ties, the majority language is predicted
- UNIV identified with hand-crafted regular expressions

Dictionary Approach (3)

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Dictionary	Accuracy (%)
BNC	80.09
SEMEVAL Twitter	77.61
LexNormList	79.86
Training Data	90.21
LexNormList+Training Data	**93.12**

All combinations were tried.

SVM without Context (1)

CENTRE FOR GLOBAL INTELLIGENT CONTENT

- Features
  - Character n-grams (G)
  - Presence in dictionary (D)
  - Binary indicators of word Length (L)
    - Split points determined by decision tree (J48) trained only with length of a word as a single feature
  - Capitalization (C)

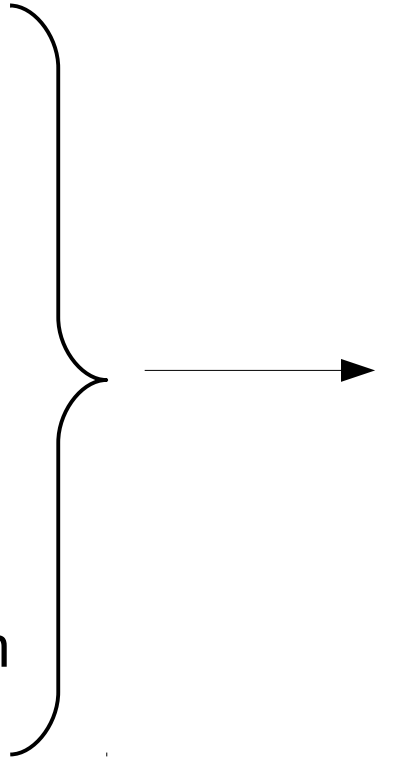
- SVM linear kernel with optimised 'C' parameter

SVM without Context (2)

- Binary indicators for length feature

J48 Pruned Tree

```
length <= 3
|   length <= 1: en
|   length > 1: bn
length > 3
|   length <= 6: bn
|   length > 6
|   |   length <= 8: bn
|   |   length > 8
|   |   |   length <= 13: en
|   |   |   length > 13: bn
```

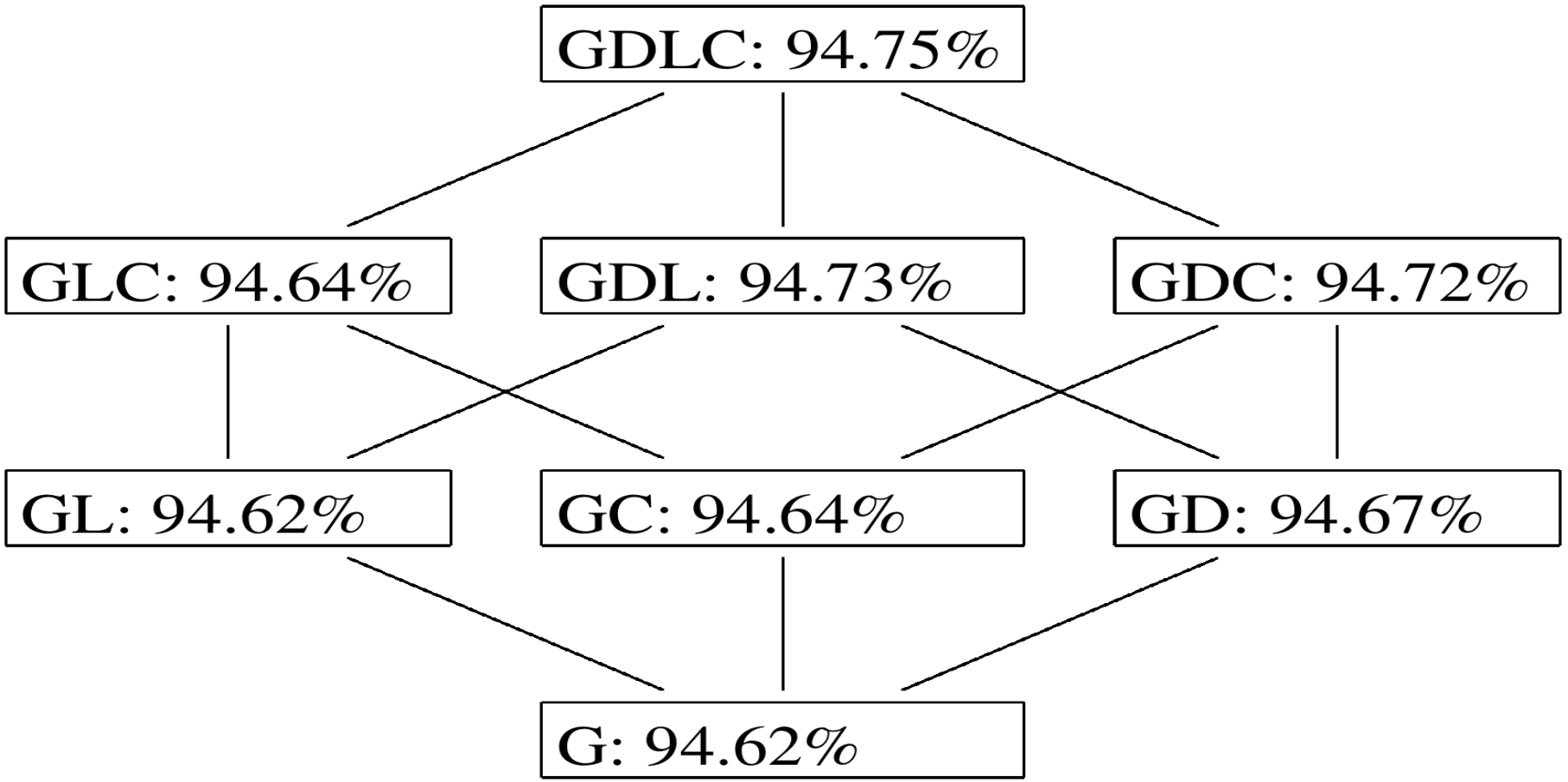


Extracted Length Features

- Is greater than 3
- Is greater than 1
- Is greater than 6
- Is greater than 8
- Is greater than 13

Encoding 6 ranges: 0-1, 2-3, 4-6, 7-8, 9-13 and 14-inf

SVM without Context (3)



SVM with Context (1)

CENTRE FOR GLOBAL INTELLIGENT CONTENT

- Features
  - Character n-grams (G)
  - Presence in dictionary (D)
  - Binary indicators of word Length (L)
  - Capitalization (C)
  - **Previous words (Pi)**
  - **Next words (Ni)**

SVM with Context (2)

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Context	Accuracy (%)
GDLC (no context)	94.75
GDLC+P2	94.66
GDLC+P1	94.55
GDLC+N1	94.53
GDLC+N2	94.37
GDLC+P1N1	**95.14**
GDLC+P2N2	94.55

CRF (1)

CENTRE FOR GLOBAL INTELLIGENT CONTENT

- Linear chain Conditional Random Field (CRF) with increasing order (0,1,2)

- Features
  - Character n-grams (G)
  - Presence in dictionary (D)
  - Word length (L)
  - Capitalisation (C)

CRF (2)

CENTRE FOR GLOBAL INTELLIGENT CONTENT

Features	Order-0	Order-1	Order-2
G	92.80	95.16	95.36
GD	93.42	95.59	95.98
GL	92.82	95.14	95.41
GDL	93.47	95.60	95.94
GC	92.07	94.60	95.05
GDC	93.47	95.62	**95.98**
GLC	92.36	94.53	95.02
GDLC	93.47	95.58	95.98

Test Set Results

CENTRE FOR GLOBAL INTELLIGENT CONTENT

- Dictionary
  - 93.64%
- SVM without context
  - 95.21%
- SVM with context
  - 95.52%
- CRF
  - 95.76%

Conclusion (1)

- Contextual clues are helpful:
  - The following example is wrongly classified by all our systems that do not use context information.
  - All context-based systems classify it correctly.

Gold data: …/univ the/en movie/en for/en which/en i/en can/en **die/en** for/en …../univ

SVM without context: …/univ the/en movie/en for/en which/en i/en can/en **die/bn** for/en …../univ

Conclusion (2)

- Character n-grams are helpful features for language identification experiments.

- Adding dictionary-based predictions as features gives a small boost to accuracy.

Another CRF Tool

We re-ran our CRF experiments with Wapiti (Lavergne et al., 2010) instead of Mallet

- 96.37% accuracy (+0.39 percentage points)

THANK YOU

SVM without Context (4)