Text Processing of Social Media I
Overview, Language Content Analysis, Language Identification

Timothy Baldwin
Talk Outline

1. Overview

2. Social Media 101
   - Introduction
   - The Appeals/Challenges of Social Media for NLP

3. Cross-comparison of the Language Content of Social Media Sources
   - Background
   - Corpora
   - Intra-Corpus Analysis
   - Inter-Corpus Analysis
   - Concluding Remarks

4. Language Identification over Social Media
   - Background
   - Datasets
   - Evaluation
Structure of the Lecture Series

- Three lectures on social media and NLP, which attempts to provide a cross-section of current research:
  - **Part 1:** overview of social media and NLP; cross-site comparison of language content; language identification over social media
  - **Part 2:** lexical normalisation; user geolocation; Twitter POS tagging
  - **Part 3:** semantic and discourse analysis of social media; user profiling; social media applications; restrictions and ethics of social media usage
- “Unashamedly” egocentric view of social media and NLP
Talk Outline

1. Overview
2. Social Media 101
   Introduction
   The Appeals/Challenges of Social Media for NLP
3. Cross-comparison of the Language Content of Social Media Sources
   Background
   Corpora
   Intra-Corpus Analysis
   Inter-Corpus Analysis
   Concluding Remarks
4. Language Identification over Social Media
   Background
   Datasets
   Evaluation
What is Social Media?

- According to Wikipedia (7/7/2014):

  Social media is the social interaction among people in which they create, share or exchange information and ideas in virtual communities and networks. Andreas Kaplan and Michael Haenlein define social media as “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content.”
What is Social Media?

- According to Wikipedia (7/7/2014):
  
  *Social media is the social interaction among people in which they create, share or exchange information and ideas in virtual communities and networks. Andreas Kaplan and Michael Haenlein define social media as “a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content.”*

**Warning**

The examples and perspective in this article deal primarily with the United States and do not represent a worldwide view of the subject.
Social Media Include ...

Social Networking sites
posts, friends/circles, “likes”, shares, events, photos, comments, geo-tags, ...

Social Media Include ...

Micro-blogs
posts, followers/followees, shares, hashtagging, geotags, ...

Source(s): http://itunes.apple.com/us/app/twitter/
Social Media Include ...

Web user forums

posts, threading, followers/followees, ...

Source(s): http://forums.cnet.com/7723-6617_102-570394/ubuntu-running-minecraft/
Social Media Include ...

- Wikis
  - posts, versioning, linking, tagging, ...

Source(s): http://en.wikipedia.org/wiki/Social_media
Source(s): http://xkcd.com/802/
Common Features of Social Media

- Posts
- Social network (explicit or implicit)
- Cross-post/user linking
- Social tagging
- Comments
- Likes/favourites/starring/...
Common Features of Social Media

- Posts
- Social network (explicit or implicit)
- Cross-post/user linking
- Social tagging
- Comments
- Likes/favourites/starring/…
- Author information, and linking to user profile features
- Aggregation/ease of access
Research on Social Media Data

- Social network analysis
- Friendship prediction/recommendation
- Search over social media data
- Social influence/dynamics
- Information dispersion
- Text mining/language technology variously ...
Properties of Social Media Data

(NLP “ideal” → actuality)

- Edited text
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
Properties of Social Media Data

(NLP “ideal” → *actuality*)

<table>
<thead>
<tr>
<th>How different?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bigram LM Perplexity:</td>
</tr>
<tr>
<td>BNC -&gt;</td>
</tr>
<tr>
<td>→ BNC</td>
</tr>
<tr>
<td>→ Twitter₁</td>
</tr>
<tr>
<td>→ Twitter₂</td>
</tr>
</tbody>
</table>
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
- Static data
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
- Streamed data
Properties of Social Media Data

(NLP “ideal” → actuality)

Challenges of Streaming Data

- throughput guarantees
- batch vs. streamed processing of data (e.g. for topic modelling)
- potential need for “incremental” models (with ability to “forget” bursty old data)

Short documents; very little context

Little language, potentially lots of other context

All over the place

What’s a sentence?

Yer what?

Anything goes — lots of languages, multilingual documents, ad hoc spelling, mix of language and markup ... language anarchy!
Properties of Social Media Data

(\text{NLP “ideal”} \rightarrow \text{actuality})

- Unedited text
- Streamed data
- Long(ish) documents; plenty of context
Properties of Social Media Data

(NLP “ideal” $\rightarrow$ actuality)

- Unedited text
- Streamed data
- Short documents; very little context
Properties of Social Media Data

(NLP “ideal” → actuality)

Document Context
Hard to adjust document-level priors when little context
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
- Streamed data
- Short documents; very little context
- All context is language context
Properties of Social Media Data

(NLP “ideal” $\rightarrow$ actuality)

- Unedited text
- Streamed data
- Short documents; very little context
- Little language, potentially lots of other context
Properties of Social Media Data

(NLP “ideal” → actuality)

Priors, priors everywhere

user priors
user-declared metadata priors
location priors
social network-based priors
hashtag priors
timezone priors
implicit social networks (retweets, user mentions, ...)

...
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
- Streamed data
- Short documents; very little context
- Little language, potentially lots of other context
- Well-defined domain/genre
Properties of Social Media Data

(NLP "ideal" \(\rightarrow\) actuality)

- *Unedited text*
- *Streamed data*
- *Short documents; very little context*
- *Little language, potentially lots of other context*
- *All over the place*
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
- Streamed data
- Short documents; very little context
- Little language, potentially lots of other context
- All over the place
- Sentence tokenisation
Properties of Social Media Data

(NLP “ideal” $\rightarrow$ actuality)

- Unedited text
- Streamed data
- Short documents; very little context
- Little language, potentially lots of other context
- All over the place
- What’s a sentence?
Properties of Social Media Data

(NLP “ideal” \(\rightarrow\) actuality)

- Unedited text
- Streamed data
- Short documents; very little context
- Little language, potentially lots of other context
- All over the place
- What’s a sentence?
- Grammaticality
Properties of Social Media Data

(NLP “ideal” $\rightarrow$ actuality)

- Unedited text
- Streamed data
- Short documents; very little context
- Little language, potentially lots of other context
- All over the place
- What’s a sentence?
- Yer what?
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
- Streamed data
- Short documents; very little context
- Little language, potentially lots of other context
- All over the place
- What’s a sentence?
- Yer what?
- Most of what glitters is English (and if your method can handle one language, it can handle ’em all)
Properties of Social Media Data

(NLP “ideal” → actuality)

- Unedited text
- Streamed data
- Short documents; very little context
- Little language, potentially lots of other context
- All over the place
- What’s a sentence?
- Yer what?
- Anything goes — lots of languages, multilingual documents, ad hoc spelling, mix of language and markup ... language anarchy!
Much of the work that is currently being carried out over social media data doesn’t make use of NLP
Observation/Questions

- Much of the work that is currently being carried out over social media data doesn’t make use of NLP
  - Are NLP methods not suited to social media analysis?
Observation/Questions

- Much of the work that is currently being carried out over social media data doesn’t make use of NLP
  - Are NLP methods not suited to social media analysis?
  - Is social media data too challenging for modern-day NLP?
Observation/Questions

- Much of the work that is currently being carried out over social media data doesn’t make use of NLP
  - Are NLP methods not suited to social media analysis?
  - Is social media data too challenging for modern-day NLP?
  - Are simple term search-based methods sufficient for social media analysis, i.e. is NLP overkill for social media?
Possible Ways Forward

- Assuming that the issue is the accuracy of NLP tools over social media data, there are two ways to proceed:
Possible Ways Forward

• Assuming that the issue is the accuracy of NLP tools over social media data, there are two ways to proceed:
  • “adapt” the data to the NLP tools through preprocessing of various forms
Possible Ways Forward

- Assuming that the issue is the accuracy of NLP tools over social media data, there are two ways to proceed:
  - “adapt” the data to the NLP tools through preprocessing of various forms
  - “adapt” the NLP tools to the data through “domain” (de-)adaptation
Talk Outline

1. Overview

2. Social Media 101
   - Introduction
   - The Appeals/Challenges of Social Media for NLP

3. Cross-comparison of the Language Content of Social Media Sources
   - Background
   - Corpora
   - Intra-Corpus Analysis
   - Inter-Corpus Analysis
   - Concluding Remarks

4. Language Identification over Social Media
   - Background
   - Datasets
   - Evaluation
Background

- Various claims made about social media text being “noisy”

Source(s): Baldwin et al. [2013]
Background

- Various claims made about social media text being “noisy”... but
  ... how “noisy” are we talking, and in what ways?

Source(s): Baldwin et al. [2013]
Various claims made about social media text being “noisy”… but

... how “noisy” are we talking, and in what ways?

... are different social media sources equally “noisy”?

Source(s): Baldwin et al. [2013]
Background

- Various claims made about social media text being “noisy”... but
  ... how “noisy” are we talking, and in what ways?
  ... are different social media sources equally “noisy”?
  ... ultimately, are the differences between social media sources all that great?

Source(s): Baldwin et al. [2013]
Outline of Approach

1. Assemble corpora across a spectrum of social media sources
   + BNC

Source(s): Baldwin et al. [2013]
Outline of Approach

1. Assemble corpora across a spectrum of social media sources
   + BNC

2. Apply a range of analyses to each individual social media source

Source(s): Baldwin et al. [2013]
Outline of Approach

1. Assemble corpora across a spectrum of social media sources + BNC
2. Apply a range of analyses to each individual social media source
3. Perform comparative analysis between different corpus pairings

Source(s): Baldwin et al. [2013]
Corpora

- Social media sources targeted in this research:
  1. **Twitter**: micro-blog posts from Twitter
  2. **Comments**: comments from YouTube
  3. **Blogs**: blog posts from Spinn3r dataset
  4. **Forums**: forum posts from popular forums

- As a balanced, non-social media counterpoint corpus:
  6. **BNC**: written portion of British National Corpus

*Source(s):* Baldwin et al. [2013]
Twitter

• 1M posts from garden hose feed of Twitter, in the form of two sub-corpora collected at different times:
  
  A Twitter-1 = 22 September 2011  
  B Twitter-2 = 22 February 2012

• Max document length = 140 characters; single author per document; no post-editing

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter-1</td>
<td>1,000,000</td>
<td>11.8±8.3</td>
</tr>
<tr>
<td>Twitter-2</td>
<td>1,000,000</td>
<td>11.6±8.1</td>
</tr>
</tbody>
</table>
Comments

- All comments associate with YouTube videos in dataset of O’Callaghan et al. [2012]
- Max document length = 500 characters; single author per document; no post-editing

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comments</td>
<td>874,772</td>
<td>15.8±18.6</td>
</tr>
</tbody>
</table>
Forums

- 1M randomly-selected posts from top-1000 vBulletin-based forums in the Big Boards forum ranking
- Max document length = site variable; single author per document; option for post-editing

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>FORUMS</td>
<td>1,000,000</td>
<td>23.2±29.3</td>
</tr>
</tbody>
</table>
BLOGS

- 1M randomly-selected documents from Spinn3r dataset (ICWSM-2011 tier-one)
- Max document length = none; single author per document; post-editing possible

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLOGS</td>
<td>1,000,000</td>
<td>147.7±339.3</td>
</tr>
</tbody>
</table>
**WIKIPEDIA**

- 200K randomly-selected documents ($\geq 500$ bytes) from English Wikipedia
- Mediawiki markup removed with `wikidump`
- Max document length = none; multiple authors per document; post-editing possible

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIKIPEDIA</td>
<td>200,000</td>
<td>281.2±363.8</td>
</tr>
</tbody>
</table>
• All documents in written portion of the British National Corpus
• Max document length = none; mostly single-author documents; post-editing possible

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Documents</th>
<th>Average words per document</th>
</tr>
</thead>
<tbody>
<tr>
<td>BNC</td>
<td>3141</td>
<td>31609.0±30424.3</td>
</tr>
</tbody>
</table>
Corpus Preprocessing

- We apply the following preprocessing to all corpora:
  1. language identification with langid.py; non-English documents filtered from corpus
  2. sentence-tokenise with tokenizer, based on findings of Read et al. [2012]
  3. tokenise and POS tag with TweetNLP 0.3
  4. remove all “non-linguistic” tokens, on basis of TweetNLP

  @helloworld Swinging with the #besties! #awesome
  ⇓

  @helloworld Swinging with the #besties! #awesome

- Extract a random sub-sample of 4K sentences

Source(s): http://www.cis.uni-muenchen.de/~wastl/misc/; Baldwin et al. [2013]
Language Mix

Combined:

Source(s): Baldwin et al. [2013]
Language Mix

Twitter:

Source(s): Baldwin et al. [2013]
Language Mix

Comments:

Source(s): Baldwin et al. [2013]
Language Mix

Forums:

Source(s): Baldwin et al. [2013]
Language Mix

Blogs:

Source(s): Baldwin et al. [2013]
Language Mix

WIKIPEDIA:

Source(s): Baldwin et al. [2013]
Language Mix

BNC:

Source(s): Baldwin et al. [2013]
Language Mix: Overall Findings

- **Twitter** most multilingual (> 50% non-EN), followed by **Comments**, **Blogs** and **Forums**
- All 97 languages modelled by `langid.py` found in **Twitter** and **Comments**

**Source(s):** Baldwin et al. [2013]
Language Mix: Overall Findings

- **Twitter** most multilingual (> 50% non-EN), followed by **Comments**, **Blogs** and **Forums**
- All 97 languages modelled by `langid.py` found in **Twitter** and **Comments**

... from here on, all analyses are based on the 4K EN sentence subset for each corpus

*Source(s):* Baldwin et al. [2013]
Lexical Analysis

- Analysis of the average word and sentence length:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Word length</th>
<th>Sentence length</th>
</tr>
</thead>
<tbody>
<tr>
<td>TWITTER-1</td>
<td>3.8±2.4</td>
<td>9.2±6.4</td>
</tr>
<tr>
<td>TWITTER-2</td>
<td>3.8±2.4</td>
<td>9.0±6.3</td>
</tr>
<tr>
<td>COMMENTS</td>
<td>3.9±3.2</td>
<td>10.5±10.1</td>
</tr>
<tr>
<td>FORUMS</td>
<td>3.8±2.3</td>
<td>14.2±12.7</td>
</tr>
<tr>
<td>BLOGS</td>
<td>4.1±2.8</td>
<td>18.5±24.8</td>
</tr>
<tr>
<td>WIKIPEDIA</td>
<td>4.5±2.8</td>
<td>21.9±16.2</td>
</tr>
<tr>
<td>BNC</td>
<td>4.3±2.8</td>
<td>19.8±14.5</td>
</tr>
</tbody>
</table>

Source(s): Baldwin et al. [2013]
Lexical Analysis

- Analysis of the rate of out-of-vocabulary words (cf. aspell):

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Word length</th>
<th>Sentence length</th>
<th>%OOV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter-1</td>
<td>3.8±2.4</td>
<td>9.2±6.4</td>
<td>24.6</td>
</tr>
<tr>
<td>Twitter-2</td>
<td>3.8±2.4</td>
<td>9.0±6.3</td>
<td>24.0</td>
</tr>
<tr>
<td>Comments</td>
<td>3.9±3.2</td>
<td>10.5±10.1</td>
<td>19.8</td>
</tr>
<tr>
<td>Forums</td>
<td>3.8±2.3</td>
<td>14.2±12.7</td>
<td>18.1</td>
</tr>
<tr>
<td>Blogs</td>
<td>4.1±2.8</td>
<td>18.5±24.8</td>
<td>20.6</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>4.5±2.8</td>
<td>21.9±16.2</td>
<td>19.0</td>
</tr>
<tr>
<td>BNC</td>
<td>4.3±2.8</td>
<td>19.8±14.5</td>
<td>16.9</td>
</tr>
</tbody>
</table>

Source(s): Baldwin et al. [2013]
**Lexical Analysis**

- Analysis of the rate of out-of-vocabulary words with lexical normalisation [Han et al., 2012]:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Word length</th>
<th>Sentence length</th>
<th>−norm</th>
<th>+norm</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Twitter-1</strong></td>
<td>3.8±2.4</td>
<td>9.2±6.4</td>
<td>24.6</td>
<td>22.5</td>
</tr>
<tr>
<td><strong>Twitter-2</strong></td>
<td>3.8±2.4</td>
<td>9.0±6.3</td>
<td>24.0</td>
<td>22.2</td>
</tr>
<tr>
<td><strong>Comments</strong></td>
<td>3.9±3.2</td>
<td>10.5±10.1</td>
<td>19.8</td>
<td>18.4</td>
</tr>
<tr>
<td><strong>Forums</strong></td>
<td>3.8±2.3</td>
<td>14.2±12.7</td>
<td>18.1</td>
<td>17.1</td>
</tr>
<tr>
<td><strong>Blogs</strong></td>
<td>4.1±2.8</td>
<td>18.5±24.8</td>
<td>20.6</td>
<td>20.3</td>
</tr>
<tr>
<td><strong>Wikipedia</strong></td>
<td>4.5±2.8</td>
<td>21.9±16.2</td>
<td>19.0</td>
<td>18.8</td>
</tr>
<tr>
<td><strong>BNC</strong></td>
<td>4.3±2.8</td>
<td>19.8±14.5</td>
<td>16.9</td>
<td>16.8</td>
</tr>
</tbody>
</table>

*Source(s):* Baldwin et al. [2013]
Lexical Analysis: Overall Findings

- Slight difference in average word length; much larger difference in average sentence length:
  \{ \text{Wikipedia, BNC, Blogs} \} > \text{Forums} > \{ \text{Comments, Twitter} \}

- With OOV\%, \text{Twitter-1/2} is the obvious outlier, although we found around half of this effect to be ameliorated through lexical normalisation

\textbf{Source(s):} Baldwin et al. [2013]
Grammaticality

- Analyse the grammaticality of sentences in the different corpora using the English Resource Grammar (ERG), judged as:
  1. **strict** (e.g. *Saarbrücken is great.*) vs. **informal** (e.g. *saarbrücken is great*)
  2. **full sentence** (e.g. *I am loving Saarbrücken*) vs. **fragment** (e.g. *Loving Saarbrücken*)

- Modified the unknown word handling in the ERG to accept POS tags from TweetNLP 0.3, and also modified the TweetNLP tokenisation slightly

- NB results not representative of “full-strength” ERG for a given data source

Source(s): Baldwin et al. [2013]
### Grammaticality Results (%)

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Parseable</th>
<th></th>
<th>Unparseable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>strict</td>
<td>informal</td>
<td></td>
</tr>
<tr>
<td></td>
<td>full frag</td>
<td>full frag</td>
<td></td>
</tr>
<tr>
<td><strong>Twitter-1</strong></td>
<td>13.8</td>
<td>23.9</td>
<td>22.2</td>
</tr>
<tr>
<td><strong>Twitter-2</strong></td>
<td>13.9</td>
<td>23.8</td>
<td>22.8</td>
</tr>
<tr>
<td><strong>Comments</strong></td>
<td>18.0</td>
<td>22.2</td>
<td>26.4</td>
</tr>
<tr>
<td><strong>Forums</strong></td>
<td>23.9</td>
<td>14.1</td>
<td>24.7</td>
</tr>
<tr>
<td><strong>Blogs</strong></td>
<td>25.6</td>
<td>17.5</td>
<td>18.8</td>
</tr>
<tr>
<td><strong>Wikipedia</strong></td>
<td>48.7</td>
<td>4.5</td>
<td>18.9</td>
</tr>
<tr>
<td><strong>BNC</strong></td>
<td>38.4</td>
<td>12.0</td>
<td>24.0</td>
</tr>
</tbody>
</table>

**Source(s):** Baldwin et al. [2013]
Grammaticality Findings

- **WIKIPEDIA** and **BNC** more grammatical than other sources; **Comments > {Blogs, Forums} > Twitter**
- More fragments in **Twitter, Comments**
- Ungrammaticality underestimated, esp. for **Blogs, Wikipedia** and **BNC**
- Preprocessing is a common cause of parser failure, with sentence tokenisation being the primary contributor
- Overall, syntactic “noise” in social media relatively minor; relative constant level of syntactic “noise” in **Twitter, Comments and Forums**

Source(s): Baldwin et al. [2013]
Corpus Similarity

- Next, we compare the (original) corpora with one another by:
  1. $\chi^2$-based corpus homogeneity [Kilgarriff, 2001]
  2. trigram language model-based perplexity

Source(s): Baldwin et al. [2013]
Corpus Similarity

- Inter-corpus similarity based on $\chi^2$:

<table>
<thead>
<tr>
<th></th>
<th>TWITTER-1</th>
<th>TWITTER-2</th>
<th>COMMENTS</th>
<th>FORUMS</th>
<th>BLOGS</th>
<th>WIKIPEDIA</th>
</tr>
</thead>
<tbody>
<tr>
<td>TWITTER-2</td>
<td>4.0</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Comments</td>
<td>63.7</td>
<td>62.4</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Forums</td>
<td>91.8</td>
<td>90.6</td>
<td>62.3</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Blogs</td>
<td>115.8</td>
<td>119.1</td>
<td>128.4</td>
<td>61.7</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Wikipedia</td>
<td>347.8</td>
<td>360.0</td>
<td>351.4</td>
<td>280.2</td>
<td>157.7</td>
<td>—</td>
</tr>
<tr>
<td>BNC</td>
<td>251.8</td>
<td>258.8</td>
<td>245.2</td>
<td>164.1</td>
<td>78.7</td>
<td>92.5</td>
</tr>
</tbody>
</table>

- Overall, this suggests the partial ordering: TWITTER-1 $\equiv$ TWITTER-2 < COMMENTS < FORUMS < BLOGS < BNC < WIKIPEDIA

Source(s): Baldwin et al. [2013]
Corpus Homogeneity

- Corpus-level homogeneity based on $\chi^2$:

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Homogeneity</th>
</tr>
</thead>
<tbody>
<tr>
<td>TWITTER-1</td>
<td>549</td>
</tr>
<tr>
<td>TWITTER-2</td>
<td>553</td>
</tr>
<tr>
<td>COMMENTS</td>
<td>613</td>
</tr>
<tr>
<td>FORUMS</td>
<td>570</td>
</tr>
<tr>
<td>BLOGS</td>
<td>716</td>
</tr>
<tr>
<td>WIKIPEDIA</td>
<td>575</td>
</tr>
<tr>
<td>BNC</td>
<td>542</td>
</tr>
</tbody>
</table>

- **BLOGS** are much more diverse in (lexical) content than other sources

Source(s): Baldwin et al. [2013]
Language Model-based Similarity
Corpus Similarity Findings

- Overall, there appears to be a partial ordering in corpus similarity: Twitter-1/2 ≡ Comments < Forums < Blogs < BNC < Wikipedia
- Internal homogeneity is broadly similar across social media sources, with Blogs as an outlier
- The “median” social media source (in terms of relative similarity to other sources) is Forums
- There is an observable difference in content between Twitter over different time periods, although it is relatively modest

Source(s): Baldwin et al. [2013]
Summary of Overall Findings

- Large-scale analysis of a range of social media sources (including BNC), in terms of language distribution, lexical analysis, grammaticality, and two measures of corpus similarity.

Source(s): Baldwin et al. [2013]
Summary of Overall Findings I

- Large-scale analysis of a range of social media sources (+ BNC), in terms of language distribution, lexical analysis, grammaticality, and two measures of corpus similarity
- How “noisy” are social media sources, and in what ways?

Source(s): Baldwin et al. [2013]
Summary of Overall Findings I

- Large-scale analysis of a range of social media sources (BNC), in terms of language distribution, lexical analysis, grammaticality, and two measures of corpus similarity
- How “noisy” are social media sources, and in what ways?
  - vastly differing levels of multilingual “noise”

Source(s): Baldwin et al. [2013]
Summary of Overall Findings I

- Large-scale analysis of a range of social media sources (+ BNC), in terms of language distribution, lexical analysis, grammaticality, and two measures of corpus similarity
- How “noisy” are social media sources, and in what ways?
  - vastly differing levels of multilingual “noise”
  - some lexical noise (esp. Twitter), but POS tagging + lexical normalisation significantly reduces it

Source(s): Baldwin et al. [2013]
Summary of Overall Findings

- Large-scale analysis of a range of social media sources (+ BNC), in terms of language distribution, lexical analysis, grammaticality, and two measures of corpus similarity.
- How “noisy” are social media sources, and in what ways?
  - vastly differing levels of multilingual “noise”
  - some lexical noise (esp. TWITTER), but POS tagging + lexical normalisation significantly reduces it
  - certainly grammatical “noise” (esp. TWITTER-1/2), but under preprocessing, the relative differences aren’t that great.

Source(s): Baldwin et al. [2013]
Summary of Overall Findings II

- Are different social media sources equally “noisy”?

Source(s): Baldwin et al. [2013]
Summary of Overall Findings II

• Are different social media sources equally “noisy”?
• no; TWITTER definitely on the noisy end of the spectrum, and WIKIPEDIA “cleaner” than BNC, but the spread is less than you might think

Source(s): Baldwin et al. [2013]
Summary of Overall Findings II

- Are different social media sources equally “noisy”? 
  - no; Twitter definitely on the noisy end of the spectrum, and Wikipedia “cleaner” than BNC, but the spread is less than you might think
- Ultimately, are the differences between social media sources all that great?

Source(s): Baldwin et al. [2013]
Summary of Overall Findings II

- Are different social media sources equally “noisy”?  
  - no; Twitter definitely on the noisy end of the spectrum, and Wikipedia “cleaner” than BNC, but the spread is less than you might think

- Ultimately, are the differences between social media sources all that great?  
  - social media spans a broad continuum from pristine to somewhat noisy, but the span is perhaps less broad than conventionally thought

Source(s): Baldwin et al. [2013]
Talk Outline

1. Overview

2. Social Media 101
   - Introduction
   - The Appeals/Challenges of Social Media for NLP

3. Cross-comparison of the Language Content of Social Media Sources
   - Background
   - Corpora
   - Intra-Corpus Analysis
   - Inter-Corpus Analysis
   - Concluding Remarks

4. Language Identification over Social Media
   - Background
   - Datasets
   - Evaluation
What is Language Identification (LangID)?

Natural Language processing (NLP) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. In theory, natural-language processing is a very attractive method of human-computer interaction.
What is Language Identification (LangID)?

RT @ThotsOnTees: Its not rocket science.....Man was designed to fail.So to those that av their trust in Man,goodluck...mine is on GOD!

実行はいつなされるんですか？ RT“@a_X_o: 制服プレイのメールの返事をすれば「いやらしすぎる」だけ帰ってきたんだけどもうなんか嫌だ”

@Luii_S2_KiSeop 哈哈哈 四次元这名号很match他xDDDDD 姐不是official kissme 么？？要怎样才能成为官方km??

#Campiglio stellata, freddo, neve dura, ma sufficiente. Rossi, Alo e Massa ok, Hayden spalla ancora immobile e dolorante.Domani #StudioSport
Why do we need LangID for Twitter?

- Twitter is highly multilingual
  - 65 Languages in 10M message sample [Bergsma et al., 2012]
  - Indigenous Tweets: 157 low-density languages (Feb 2014)
- NLP often monolingual
- Keyword-based approaches have shortcomings
  - poor recall
  - cognates/false friends
- Twitter added language predictions in March 2013
  - spoiler: it is not perfect!
  - older crawls?

Source(s): Lui and Baldwin [2014]
Challenges in LangID on Social Media

- **Short message length**: individual documents are generally short

*Source(s):* Lui and Baldwin [2014]
Challenges in LangID on Social Media

- **Short message length**: individual documents are generally short
- **Variety of registers, domains, ...**: there is a lot of variety in the content of documents

**Source(s)**: Lui and Baldwin [2014]
Challenges in LangID on Social Media

- **Short message length:** individual documents are generally short
- **Variety of registers, domains, ...:** there is a lot of variety in the content of documents
- **Lexical variation:** there is a lot of fluidity in how a given word is spelled

Source(s): Lui and Baldwin [2014]
Challenges in LangID on Social Media

- **Short message length:** individual documents are generally short
- **Variety of registers, domains, ...:** there is a lot of variety in the content of documents
- **Lexical variation:** there is a lot of fluidity in how a given word is spelled
- **Linguistic diversity:** a rich mix of languages can be found on social media sites, often with different distributions, and with no “closed-world” guarantee

*Source(s):* Lui and Baldwin [2014]
Challenges in LangID on Social Media

- **Short message length:** individual documents are generally short
- **Variety of registers, domains, ...:** there is a lot of variety in the content of documents
- **Lexical variation:** there is a lot of fluidity in how a given word is spelled
- **Linguistic diversity:** a rich mix of languages can be found on social media sites, often with different distributions, and with no “closed-world” guarantee
- **Limited labelled corpora:** language-labelled corpora of social media data are few and far between

Source(s): Lui and Baldwin [2014]
Another Complication

Source Domain

⇒

Target Domain

SUPERVISED LEARNING

INDUCTIVE TRANSFER LEARNING

TRANSDUCTIVE TRANSFER LEARNING
Another Complication II

- LangID Accuracy drops considerably when moving from a supervised to an inductive transfer learning setting

Source(s): Lui and Baldwin [2012]
Selecting Features with Cross-domain Utility I

- **Observation:** per-domain vs. per-language information gain of byte $n$-grams ($1 \leq n \leq 4$):
Selecting Features with Cross-domain Utility II

• Solution:
  
  • pre-filter the feature set according to DF (the number of documents containing a given feature), based on the observation that low DF $\rightarrow$ low IG

  $$\mathcal{LD}^{bin}(t|l) = \mathcal{IG}_{language}^{bin}(t|l) - \mathcal{IG}_{domain}(t)$$

  
  • calculate the information gain for each of the top-$N$ features $t$ relative to a given language $l$, using a simple subtractive method to compensate for domain effects:

  Source(s): Lui and Baldwin [2011]
# Benchmarking LangID over Twitter

<table>
<thead>
<tr>
<th>Tool</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>langid.py</code></td>
<td>Lui and Baldwin [2012]</td>
</tr>
<tr>
<td>ChromeCLD</td>
<td>McCandless [2010]</td>
</tr>
<tr>
<td>LangDetect</td>
<td>Nakatani [2010]</td>
</tr>
<tr>
<td>LDIG</td>
<td>Nakatani [2012]</td>
</tr>
<tr>
<td>whatlang</td>
<td>Brown [2013]</td>
</tr>
<tr>
<td>YALI</td>
<td>Majliš [2012]</td>
</tr>
<tr>
<td>TextCat</td>
<td>Scheelen [2003]</td>
</tr>
<tr>
<td>MSR-LID</td>
<td>Goldszmidt et al. [2013]</td>
</tr>
</tbody>
</table>
## Manually-labelled Datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Message Count</th>
<th>Languages (ISO639-1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benelearn</td>
<td>9066</td>
<td>(6) de es en fr it nl</td>
</tr>
<tr>
<td>SCarter</td>
<td>5000</td>
<td>(5) de es en fr nl</td>
</tr>
<tr>
<td>Bergsma</td>
<td>13190</td>
<td>(9) ar bg fa hi mr ne ru uk ur</td>
</tr>
<tr>
<td>ZhEnJa</td>
<td>3016</td>
<td>(3) en ja zh</td>
</tr>
</tbody>
</table>
Bergsma et al. [2012]

- 9 languages
- 3 non-Latin scripts (Arabic, Cyrillic, Devanagari)
- Off-the-shelf results (% accuracy):

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>langid.py</td>
<td>91.4</td>
<td>78.4</td>
<td>88.8</td>
</tr>
<tr>
<td>ChromeCLD</td>
<td>90.5</td>
<td>N/A</td>
<td>91.4</td>
</tr>
</tbody>
</table>
**ZhENJA dataset**

- Randomly sampled 5000 messages from a 24h period
- Each message annotated by speakers of:
  - Mandarin Chinese (zh)
  - English (en)
  - Japanese (ja)
- Each message annotated by 3 speakers of each language
- Total 3016 messages labeled:
  - zh: 16 (0.3%)
  - en: 1953 (39.1%)
  - ja: 1047 (20.9%)

*Source(s):* Lui and Baldwin [2014]
Evaluating Off-the-Shelf LangID Systems on Twitter
Which LangID System to Use?

- Used the 4 existing datasets to evaluate all 8 systems
- Evaluated 3-system, 5-system and 7-system majority vote
- Voting compensates for individual systems’ weakness on Bergsma and ZhEnJa
- 3-system vote between `langid.py`, ChromeCLD and LangDetect is a good choice
**TwitUser Dataset**

- Active: 5 msgs/day on 7 different days (31 days of data)
- 65 languages
- Top: en (50.6%) ja (14.1%) pt (13.0%)
- Randomly selected up to 100 users per language
- Pool of 26011 msgs from 2914 users
TwitUser: Off-the-Shelf Results
**TwitUser: Off-the-Shelf Results**

![Diagram showing comparison of various text processing tools for language identification.](image-url)
Custom Tweaks for Social Media Processing

- **Tweak:** “clean” the text to pre-remove URLs, user mentions, hashtags and smilies [Tromp and Pechenizkiy, 2011]

- **Motivation:** “metadata” in the tweet is often in ASCII and misleading in terms of language content

- **Practicum:**
  - very easy to implement
  - no difference on Twitter-specific systems (MSR-LID, LDIG)
  - small improvement on other systems (< 2%)
Custom Tweaks for Social Media Processing II

- **Tweak:** bootstrapping, i.e. train the classifier on automatically-labelled tweets [Goldszmidt et al., 2013]
- **Motivation:** data is cheap but labels are expensive; large volumes of slightly noisy training data > smaller volumes of pristine training data
- **Practicum:**
  - requires re-training classifier
  - tested with `LangDetect`, `TextCat`, `langid.py`
  - used 2 methods to construct bootstrap collection (direct `LangID`, user prior)
  - not consistently better than off-the-shelf
Custom Tweaks for Social Media Processing III

- **Tweak:** learn LangID priors of different types (user, link, mention, hashtag, thread), and incorporate these into the classifier [Carter et al., 2013, Bontcheva et al., 2013]

- **Motivation:** individual users will tend to use only a small set of languages; these are strong predictors of what languages they are likely to tweet in in the future

- **Practicum:**
  - requires processing a massive background collection, and storing a large set of user priors
  - Carter et al. [2013] and Bontcheva et al. [2013] report positive results
Twitter API predictions

- While it is against Twitter’s Terms of Service to benchmark against the Twitter LangID predictions, we can make some observations:
  - 25% of tweets in TwitUser are not LangID tagged
  - Some obvious language gaps, e.g. all Romanian messages are identified as Italian
  - Twitter’s predictions are:
    - not perfect
    - not substantially better than off-the-shelf
Conclusion

- Importance of “domain deadaptation” in LangID, because of: (a) lack of large-scale training corpora; and (b) susceptibility of LangID systems to negative transfer
- LangID over Twitter far from perfect (state-of-the-art = 0.89 F-Score), and Twitter API certainly not perfect
- Scope for improvement in accuracy through use of priors of various types (going beyond the individual message)
- Further research needed to broaden scope and increase accuracy
References


References II


References III


