Automatic Information Extraction from Unstructured Text – practices and case studies

SHUBHANSHU MISHRA
ISCHOOL AT ILLINOIS
@THESHUBHANSHU
## Structured v/s Unstructured Text

<table>
<thead>
<tr>
<th>Structured Text</th>
<th>Un-Structured Text</th>
</tr>
</thead>
</table>
| **Persons:** Rafa → wiki(Rafael_Nadal), Fed → wiki(Roger Federer) | **Rafa and Fed take on each other on May 2\textsuperscript{nd}**  
**USOpen**  
**GoRafa.**  
*Timestamp:* April 31\textsuperscript{st}, 2017 |
| **Event:** USOpen → wiki(US Open 2017) |  |
| **Time:** May 2\textsuperscript{nd}, 2017 |  |
| **Sentiment:** Support towards Rafa |  |
| • Rarely available  
• Unique representation  
• Easy to parse and load in database  
• Easy to answer questions on  
• Easy to search |  |
|  | • Abundantly available  
• Multiple representations  
• Easy to parse if data is proper English (e.g. newswire corpus)  
• Harder to parse when the language used is different |
Information extraction issues

Source: https://twitter.com/BillSimons1/status/854713304949825540

Source: http://corenlp.run/

Text to annotate:

The 10s Gods Must B Crazy/Could we have a10s wrld where #Federer, 35, wins 2 huge tourneys & King of Clay, Senor #Nadal, loses in 2nd rd.

Annotations:

Part-of-Speech:

Named Entity Recognition:

Wikitdict Entities:

Sentiment:

- Is it even right to assign a polarity to the overall text?
- What does positive or negative sentiment mean? (taken from ML lingo)
- Shouldn’t sentiment convey action or internal state of—either the reader or the author?
Traditional NLP tools break on Social Media Data

Part-of-Speech:
- Emoticon lost

Named Entity Recognition:
- No entities captured

Sentiment:
- Sentiment lost

Source: http://corenlp.run/
“Our field is the **domain science of language technology**; it’s **not** about the best method of machine learning—the central issue remains the **domain problems**. The domain problems **will not go away**.”

--- **Christopher Manning** (ACL President 2015)

## Solutions for NLP in Social Media Text

### Sentiment Analysis:

- Define **actionable labels** for sentiment to capture if the author is:
  - **Enthusiastic or passive** in their expression
  - **Supportive or non-supportive** of the topic
  - Make task specific labels – don’t just stick to positive or negative

- Utilize the **context of the text** for making prediction:
  - Who is the **user**? – number of followers, friends, statuses, etc.
  - What **metadata** is associated with the tweet? – retweets, favorites, etc.

### Named Entity Recognition and Part of Speech Tagging:

- Allow models to:
  - Change with **time**
  - Adapt to your testing data

- This can be achieve via:
  - Updating the **vocabulary** of the model
  - Using **online or semi-supervised learning** models to learn from unlabeled and newer data
Enthusiasm and Support


-“All the best to the retired players suffering from CTE. Spread the word so we can make the game safer.”

-“New LGBT Research Study on same sex weddings [link]”

-“Just watched cyberbully-- it’s annoying. Why would she kill herself? It’s not worth it. Life is shit so deal with it :P”

-“Just watched cyberbully-- it’s annoying. Why would she kill herself? It’s not worth it. Life is shit so deal with it :P”
Actionable labels:
Tuition-Free Higher Education Policy Debate on Facebook

Opinion on subject
• **For:** Tax money used for a good long-term investment, finally.
• **Against:** Free? It’s not free-someone’s gotta pay.

Tone of the conversation
• **Civil:** Nothing is free Mr. President.
• **Non-Civil:** NOTHING IS FREE!!!!!! WAKE UP!!

Relatedness to the topic
• **On-Topic:** I hope they pay for it with cuts in wasteful spending, not more taxes.
• **Off-Topic:** There’s ALWAYS a catch, just like with “You can keep your insurance”

### Use context

**User attributes: Gender, Ethnicity, etc.**

<table>
<thead>
<tr>
<th><strong>Gender</strong></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>- Extracted via US SSN data</td>
<td></td>
</tr>
<tr>
<td>- Can also use existing image based classifiers</td>
<td></td>
</tr>
</tbody>
</table>

**Ethnicity**


**Race**

- Manually annotated using profile inspection

**Political Leaning**

- Manually annotated using profile inspection (if available)

---

Adaptive models (Sentiment): Vocabulary Update & Online Learning


Download tool from: https://github.com/uiuc-ischool-scanr/SAIL
Adaptive models (NER): Vocabulary Update & Semi-supervised Learning

Unlabeled Data

Word Clusters

Word Embedding

Random Feature Dropout

Global Features

Feature Extractors

Gazetteers

Regex

Labeled Data

Wiki Data

Geonames

Discogs

Labels in BIEOU format

Original Model: **46.4% F1**
Semi-supervised mode: **47.3% F1**


Download from: [https://github.com/napsternxg/TwitterNER](https://github.com/napsternxg/TwitterNER)
## Open challenges

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not enough labeled data for social media IE tasks</td>
<td></td>
</tr>
<tr>
<td>Not enough pre-trained models</td>
<td></td>
</tr>
<tr>
<td>Rapidly evolving contents</td>
<td></td>
</tr>
<tr>
<td>Privacy and TOS issues – data available today might be missing tomorrow</td>
<td></td>
</tr>
</tbody>
</table>
Key Takeaways

Don’t apply existing tools blindly
Understand your labels and their relation with the data
Use the context
Allow models to evolve (be aware if they don’t)
Questions?

Contact:
Shubhanshu Mishra
http://shubhanshu.com
@TheShubhanshu