Making natural language processing robust to sociolinguistic variation

Jacob Eisenstein
@jacobbeisenstein

Georgia Institute of Technology

September 9, 2017
Machine reading
From text to structured representations.
Machine reading
From text to structured representations.

Annotate and train
Machine reading
From text to structured representations.

New domains of digitized texts offer opportunities as well as challenges.
Language data then and now

Then: news text, small set of authors, professionally edited, fixed style
Language data then and now

**Then**: news text, small set of authors, professionally edited, fixed style

**Now**: open domain, everyone is an author, unedited, many styles
Social media has forced NLP to confront the challenge of missing social context (Eisenstein, 2013): (Gimpel et al., 2011) (Ritter et al., 2011) (Foster et al., 2011)
Social media has forced NLP to confront the challenge of missing social context (Eisenstein, 2013):

- tacit assumptions about audience knowledge
- language variation across social groups

(Gimpel et al., 2011) (Ritter et al., 2011) (Foster et al., 2011)
Social media has forced NLP to confront the challenge of missing social context (Eisenstein, 2013):

- tacit assumptions about audience knowledge
- language variation across social groups

(Gimpel et al., 2011)
(Ritter et al., 2011)
(Foster et al., 2011)
an absolutely perfect response by the warriors
an absolutely perfect response by the warriors
Finding tacit context in the social network

- Social media texts lack context, because it is implicit between the writer and the reader.

- **Homophily**: socially connected individuals tend to share traits.
Assortativity of entity references

![Graph showing assortativity of entity references for different types of mentions and retweets.](image)
an absolutely perfect response by the warriors
an absolutely perfect response by the warriors
The return of Clutch Dirk Nowitzki is one of the more exciting, unexpected developments in an already bonkers NBA season

an absolutely perfect response by the warriors
We project embeddings for entities, words, and authors into a shared semantic space.

Inner products in this space indicate compatibility.
Socially-Infused Entity Linking

\[ s(x, y, u) = g_1(x, y_t, t) + g_2(x, y_t, u, t) \]
Socially-Infused Entity Linking

\[ s(x, y, u) = g_1(x, y_t, t) + g_2(x, y_t, u, t) \]

tweet \rightarrow author
entity assignments
Socially-Infused Entity Linking

\[ s(x, y, u) = g_1(x, y_t, t) + g_2(x, y_t, u, t) \]

- \( g_1 \) is employed to model surface features.
Socially-Infused Entity Linking

\[ s(\mathbf{x}, \mathbf{y}, \mathbf{u}) = g_1(\mathbf{x}, y_t, t) + g_2(\mathbf{x}, y_t, u, t) \]

tweet \quad \downarrow \quad \text{author}
entity assignments

- \( g_1 \) is employed to model surface features.
- \( g_2 \) is used to capture two assumptions:
  - Entity homophily
  - Semantically related mentions tend to refer similar entities
Socially-Infused Entity Linking

\[ g_1(x, y_t, t; \Theta_1) = \beta^\top \tanh(W\phi(x, y_t, t) + b) + b \]

\[ g_2(x, y_t, u, t; \Theta_2) = (v_u^T W^{(u,e)} v_{y_t}^{(e)} + v_t^T W^{(m,e)} v_{y_t}^{(e)}) \]

(author embedding) \quad (mention embedding) \quad (entity embedding)
Socially-Infused Entity Linking

\[
g_1(x, y_t, t; \Theta_1) = \beta^\top \tanh(W \phi(x, y_t, t) + b) + b
\]

\[
g_2(x, y_t, u, t; \Theta_2) = v_u^u W^{u,e} v_{y_t}^{(e)} + v_t^m W^{m,e} v_{y_t}^{(e)}
\]
Learning

\[ L(\Theta) = \max_{y \in Y_x} (\Delta(y, y^*) + s(x, y, u)) - s(x, y^*, u) \]
Learning

\[ L(\Theta) = \max_{y \in \mathcal{Y}_x} (\Delta(y, y^*) + s(x, y, u)) - s(x, y^*, u) \]

- Loss-augmented inference:

\[ \hat{y} = \arg\max_{y \in \mathcal{Y}_x} (\Delta(y, y^*) + s(x, y, u)) \]
Learning

\[ L(\Theta) = \max_{y \in \mathcal{Y}_x} (\Delta(y, y^*) + s(x, y, u)) - s(x, y^*, u) \]

- Loss-augmented inference: **Hamming loss**

\[ \hat{y} = \arg \max_{y \in \mathcal{Y}_x} (\Delta(y, y^*) + s(x, y, u)) \]
\[ L(\Theta) = \max_{y \in \mathcal{Y}_x} \left( \Delta(y, y^*) + s(x, y, u) \right) - s(x, y^*, u) \]

- Loss-augmented inference: \textit{hamming loss}
  \[ \hat{y} = \arg \max_{y \in \mathcal{Y}_x} \left( \Delta(y, y^*) + s(x, y, u) \right) \]

- Optimization: stochastic gradient descent
Inference

- **Non-overlapping structure**

Tweet: thought u were a Red Sox fan

Overlapped mentions: ‘Red’ ‘Sox’ ‘Red Sox’

Entities: Nil

In order to link ‘Red Sox’ to a real entity, ‘Red’ and ‘Sox’ should be linked to Nil.
- Structure prediction improves accuracy.
- Social context yields further improvements.
- S-MART is the prior state-of-the-art (Yang & Chang, 2015).
Social media has forced NLP to confront the challenge of missing social context (Eisenstein, 2013):

- tacit assumptions about audience knowledge
- language variation across social groups

(Gimpel et al., 2011)
(Ritter et al., 2011)
(Foster et al., 2011)
Social media has forced NLP to confront the challenge of missing social context (Eisenstein, 2013):

- tacit assumptions about audience knowledge
- language variation across social groups

(Gimpel et al., 2011)
(Ritter et al., 2011)
(Foster et al., 2011)
Language variation: a challenge for NLP

“I would like to believe he’s sick rather than just mean and evil.”

(Yang & Eisenstein, 2017)
Language variation: a challenge for NLP

“I would like to believe he’s **sick** rather than just mean and evil.”

“You could’ve been getting down to this **sick** beat.”

(Yang & Eisenstein, 2017)
Personalization by ensemble

- **Goal:** personalized conditional likelihood, \( P(y \mid x, a) \), where \( a \) is the author.
- **Problem:** We have labeled examples for only a few authors.
Personalization by ensemble

- **Goal:** personalized conditional likelihood, \( P(y | x, a) \), where \( a \) is the author.
- **Problem:** We have labeled examples for only a few authors.
- **Personalization ensemble**

\[
P(y | x, a) = \sum_k P_k(y | x) \pi_a(k)
\]

- \( P_k(y | x) \) is a basis model
- \( \pi_a(\cdot) \) are the ensemble weights for author \( a \)
Homophily to the rescue?

Labeled data

Unlabeled data

Are language styles **assortative** on the social network?
Evidence for linguistic homophily

Pilot study: is classifier accuracy assortative on the Twitter social network?

$$assort(G) = \frac{1}{|G|} \sum_{(i,j) \in G} \delta(y_i = \hat{y}_i)\delta(y_j = \hat{y}_j)$$

$$+ \delta(y_i \neq \hat{y}_i)\delta(y_j \neq \hat{y}_j)$$
Evidence for linguistic homophily

Pilot study: is classifier accuracy **assortative** on the Twitter social network?

\[
\text{assort}(G) = \frac{1}{\#|G|} \sum_{(i,j) \in G} \delta(y_i = \hat{y}_i)\delta(y_j = \hat{y}_j) \\
+ \delta(y_i \neq \hat{y}_i)\delta(y_j \neq \hat{y}_j)
\]
Network-driven personalization

- For each author, estimate a **node embedding** $e_a$ (Tang et al., 2015).
- Nodes who share neighbors get similar embeddings.

\[ \pi_a = \text{SoftMax}(f(e_a)) \]

\[ P(y \mid x, a) = \sum_{k=1}^{K} P_k(y \mid x) \pi_a(k) \]
Results

Improvements over ConvNet baseline:
- +2.8% on Twitter Sentiment Analysis
- +2.7% on Ciao Product Reviews

NLSE is prior state-of-the-art (Astudillo et al., 2015).
## Variable sentiment words

<table>
<thead>
<tr>
<th>More positive</th>
<th>More negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 banging loss fever broken <strong>fucking</strong></td>
<td><strong>dear like god yeah wow</strong></td>
</tr>
<tr>
<td>2 chilling cold ill sick suck</td>
<td>satisfy trust wealth strong <strong>lmao</strong></td>
</tr>
<tr>
<td>3 <strong>ass damn piss bitch shit</strong></td>
<td>talent honestly voting win clever</td>
</tr>
<tr>
<td>4 insane bawling fever weird cry</td>
<td><strong>lmao super lol haha hahaha</strong></td>
</tr>
<tr>
<td>5 ruin silly bad boring dreadful</td>
<td><strong>lovatics wish beliebers ariana-tors kendall</strong></td>
</tr>
</tbody>
</table>
Summary

**Robustness** is a key challenge for making NLP effective on social media data:

- Tacit assumptions about shared knowledge; language variation
- Social metadata gives NLP systems the flexibility to handle each author differently.

- Word embeddings for unseen words (Pinter et al., 2017)
- Lexicon-based supervision (Eisenstein, 2017)
- Applications to finding rare events in electronic health records (ongoing work with Jimeng Sun)
Robustness is a key challenge for making NLP effective on social media data:

- Tacit assumptions about shared knowledge; language variation
- Social metadata gives NLP systems the flexibility to handle each author differently.

The long tail of rare events is the other big challenge.

- Word embeddings for unseen words (Pinter et al., 2017)
- Lexicon-based supervision (Eisenstein, 2017)
- Applications to finding rare events in electronic health records (ongoing work with Jimeng Sun)
Acknowledgments

- Students and collaborators:
  - Yi Yang (GT → Bloomberg)
  - Mingwei Chang (Google Research)
  - See https://gtnlp.wordpress.com/ for more!

- Funding: National Science Foundation, National Institutes for Health, Georgia Tech


