Making natural language processing robust to sociolinguistic variation

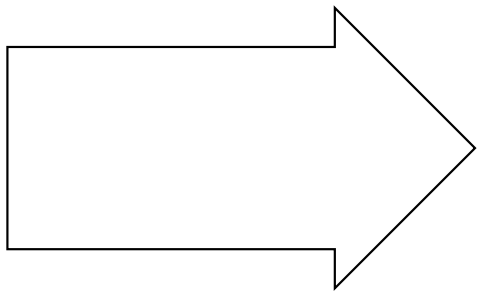
Jacob Eisenstein
@jacobeisenstein

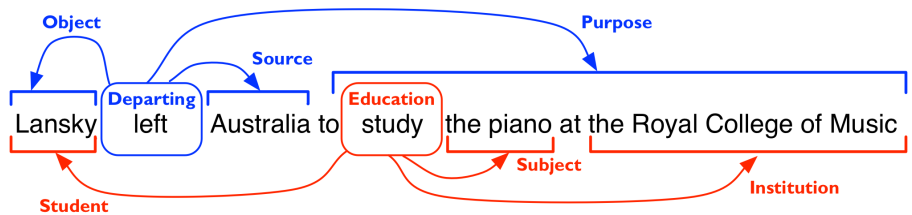
Georgia Institute of Technology

September 9, 2017

Machine reading

From text to structured representations.

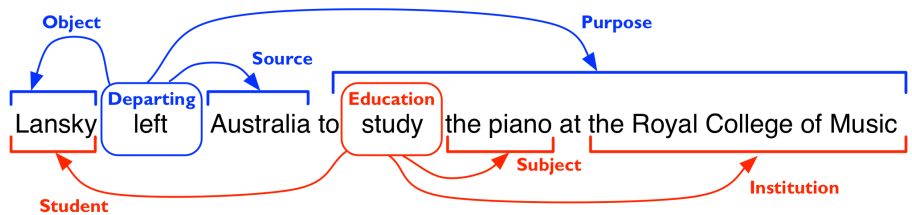




Machine reading

From text to structured representations.

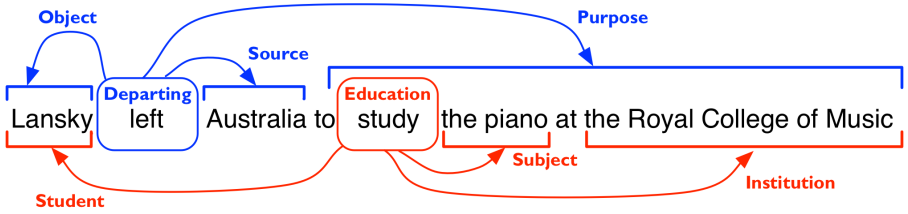
Annotate and train



Machine reading

From text to structured representations.

Annotate and train



New domains of digitized texts offer opportunities as well as challenges.

Language data then and now

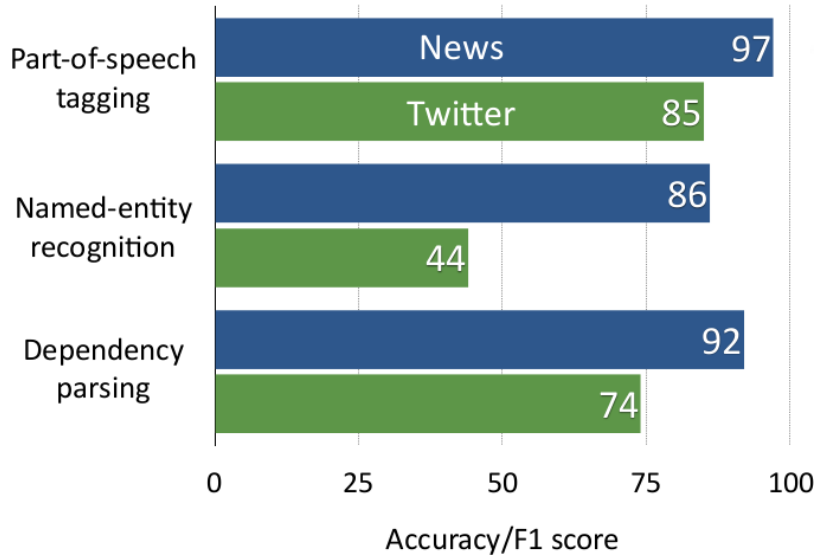
Then: news text, small set of authors, professionally edited, fixed style

Language data then and now

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

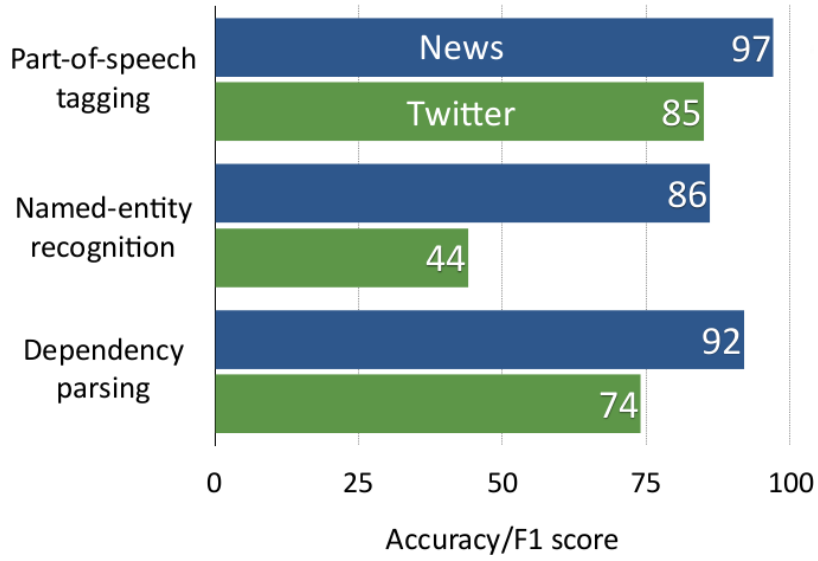
Now: open domain, everyone is an author, unedited, many styles

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)

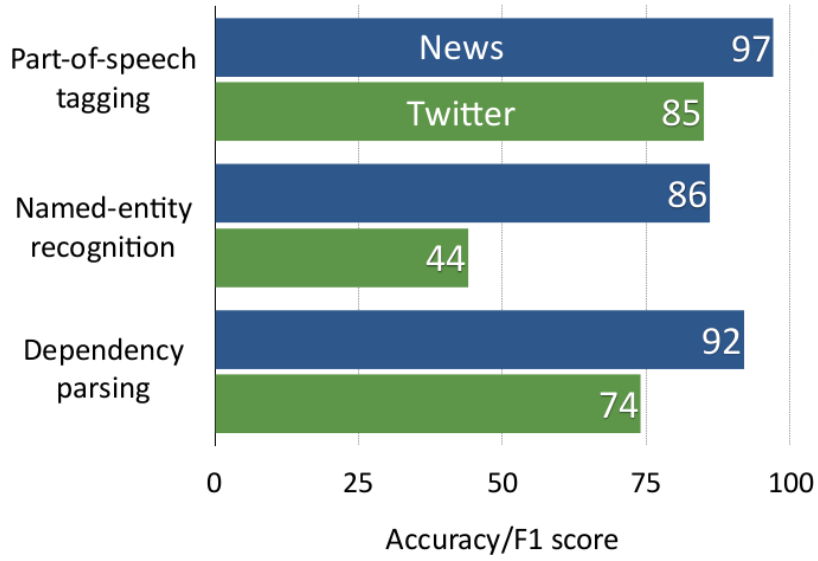
Social media has forced NLP to confront the challenge of missing social context (Eisenstein, 2013):



- 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)

Shea Serrano
@SheaSerrano

Follow

an absolutely perfect response by the warriors

Shea Serrano
@SheaSerrano

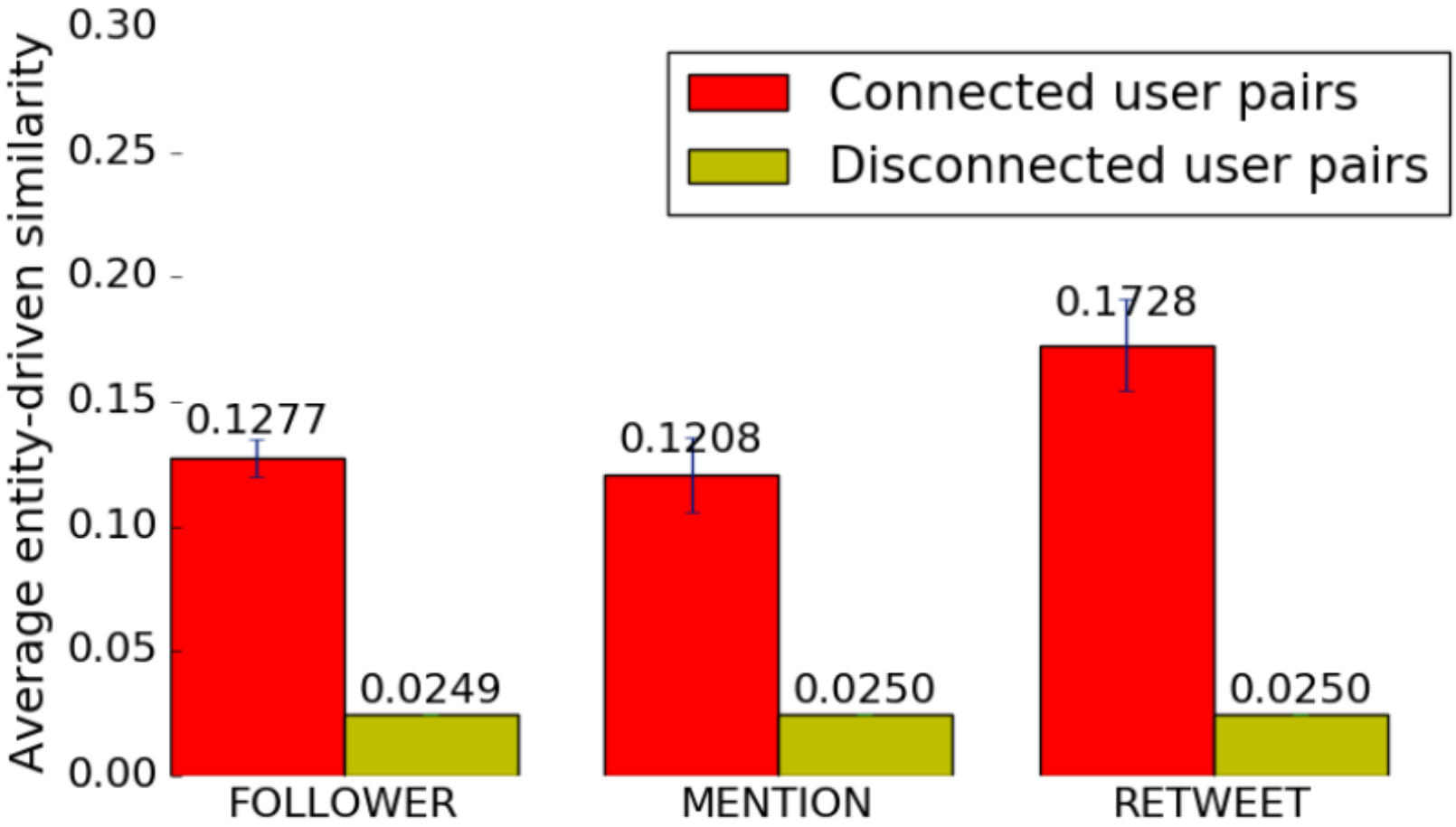
Follow

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.

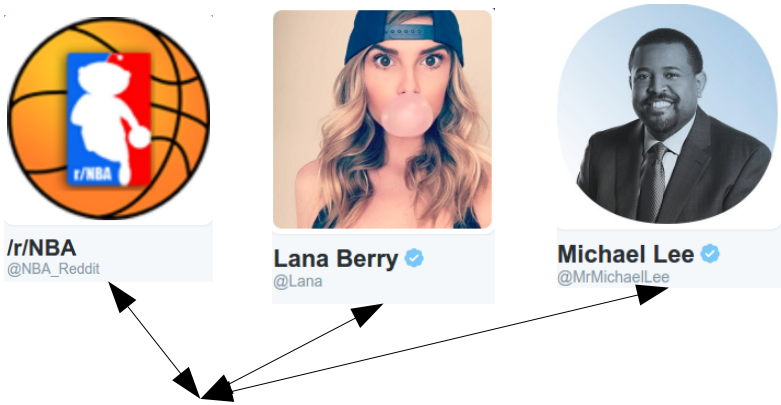
Assortativity of entity references



Shea Serrano
@SheaSerrano

Follow

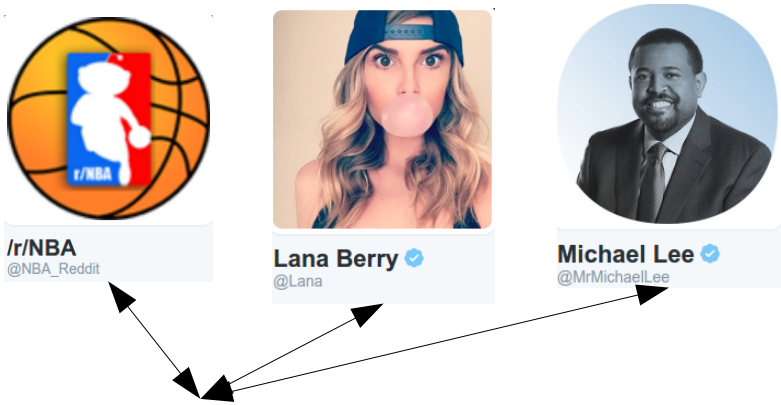
an absolutely perfect response by the warriors



Shea Serrano
@SheaSerrano

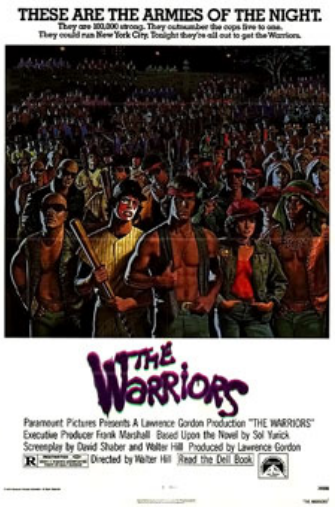
Follow

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

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

Socially-Infused Entity Linking

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

tweet

entity assignments

author

Socially-Infused Entity Linking

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

tweet
entity assignments
author

- g_1 is employed to model surface features.

Socially-Infused Entity Linking

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

tweet
entity assignments
author

- 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

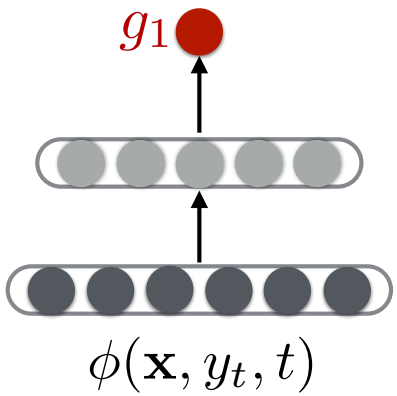
$$g_1(\mathbf{x}, y_t, t; \Theta_1) = \boldsymbol{\beta}^\top \tanh(\mathbf{W}\phi(\mathbf{x}, y_t, t) + \mathbf{b}) + b$$

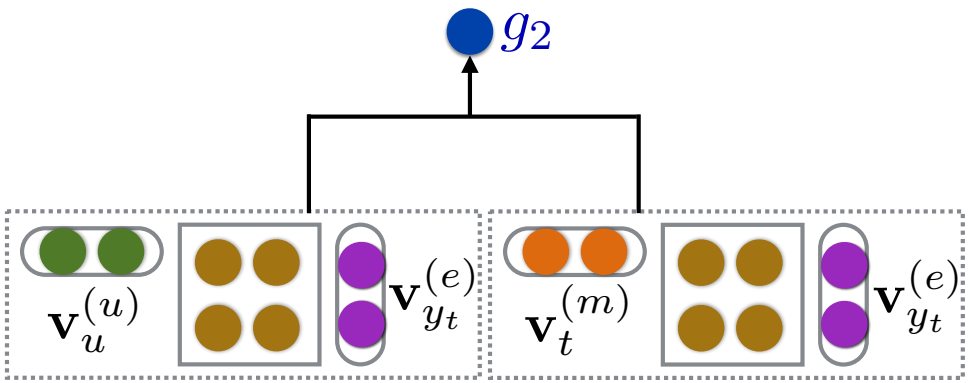
$$g_2(\mathbf{x}, y_t, u, t; \Theta_2) = \mathbf{v}_u^{(u)\top} \mathbf{W}^{(u,e)} \mathbf{v}_{y_t}^{(e)} + \mathbf{v}_t^{(m)\top} \mathbf{W}^{(m,e)} \mathbf{v}_{y_t}^{(e)}$$

author embedding

mention embedding

entity embedding





Socially-Infused Entity Linking

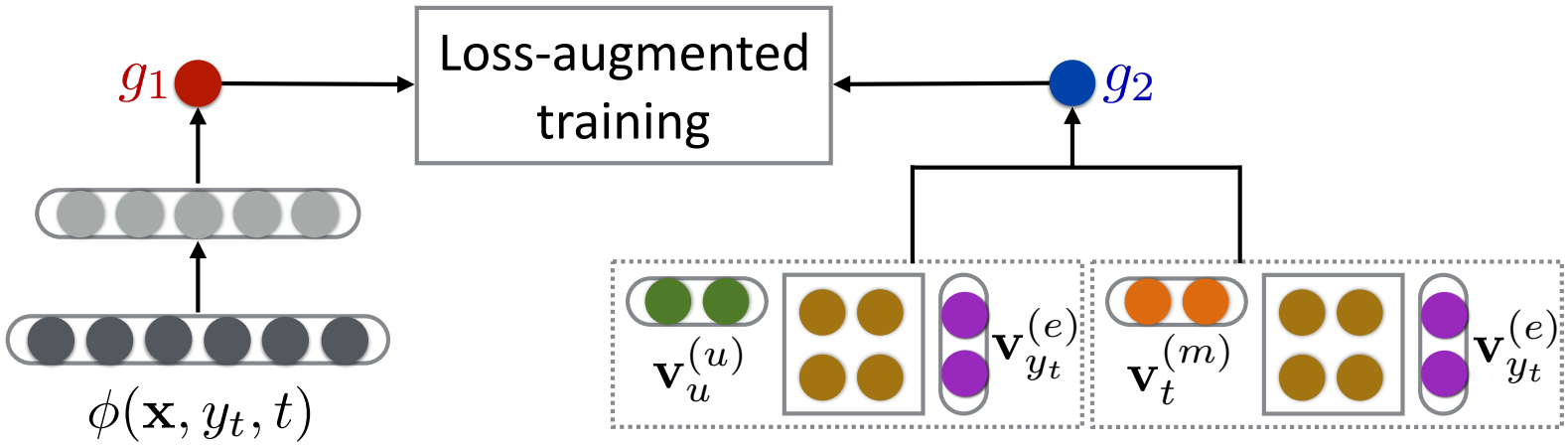
$$g_1(\mathbf{x}, y_t, t; \Theta_1) = \boldsymbol{\beta}^\top \tanh(\mathbf{W}\phi(\mathbf{x}, y_t, t) + \mathbf{b}) + b$$

$$g_2(\mathbf{x}, y_t, u, t; \Theta_2) = \mathbf{v}_u^{(u)\top} \mathbf{W}^{(u,e)} \mathbf{v}_{y_t}^{(e)} + \mathbf{v}_t^{(m)\top} \mathbf{W}^{(m,e)} \mathbf{v}_{y_t}^{(e)}$$

author embedding

mention embedding

entity embedding



Learning

$$L(\Theta) = \max_{\mathbf{y} \in \mathcal{Y}_{\mathbf{x}}} (\Delta(\mathbf{y}, \mathbf{y}^*) + s(\mathbf{x}, \mathbf{y}, u)) - s(\mathbf{x}, \mathbf{y}^*, u)$$

Learning

$$L(\Theta) = \max_{\mathbf{y} \in \mathcal{Y}_{\mathbf{x}}} (\Delta(\mathbf{y}, \mathbf{y}^*) + s(\mathbf{x}, \mathbf{y}, u)) - s(\mathbf{x}, \mathbf{y}^*, u)$$

- Loss-augmented inference:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}_{\mathbf{x}}} (\Delta(\mathbf{y}, \mathbf{y}^*) + s(\mathbf{x}, \mathbf{y}, u))$$

Learning

$$L(\Theta) = \max_{\mathbf{y} \in \mathcal{Y}_\mathbf{x}} (\Delta(\mathbf{y}, \mathbf{y}^*) + s(\mathbf{x}, \mathbf{y}, u)) - s(\mathbf{x}, \mathbf{y}^*, u)$$

- Loss-augmented inference:

hamming loss

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}_\mathbf{x}} (\Delta(\mathbf{y}, \mathbf{y}^*) + s(\mathbf{x}, \mathbf{y}, u))$$

Learning

$$L(\Theta) = \max_{\mathbf{y} \in \mathcal{Y}_\mathbf{x}} (\Delta(\mathbf{y}, \mathbf{y}^*) + s(\mathbf{x}, \mathbf{y}, u)) - s(\mathbf{x}, \mathbf{y}^*, u)$$

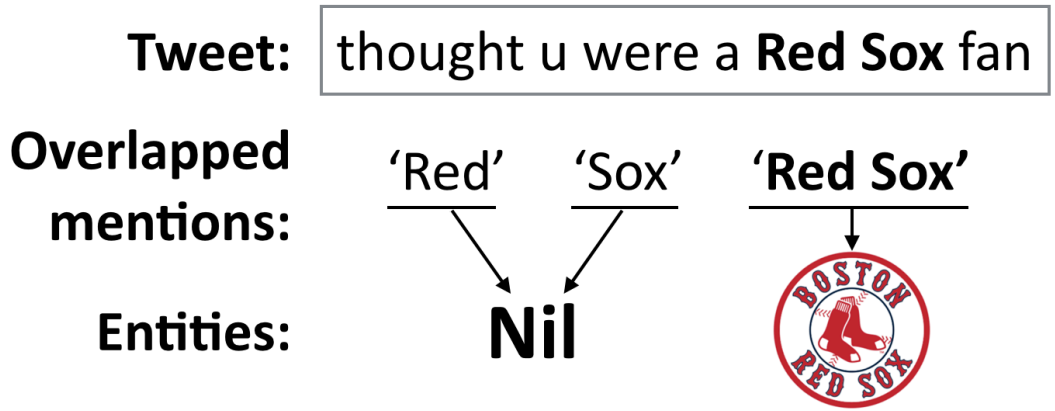
- Loss-augmented inference: hamming loss

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y} \in \mathcal{Y}_\mathbf{x}} (\Delta(\mathbf{y}, \mathbf{y}^*) + s(\mathbf{x}, \mathbf{y}, u))$$

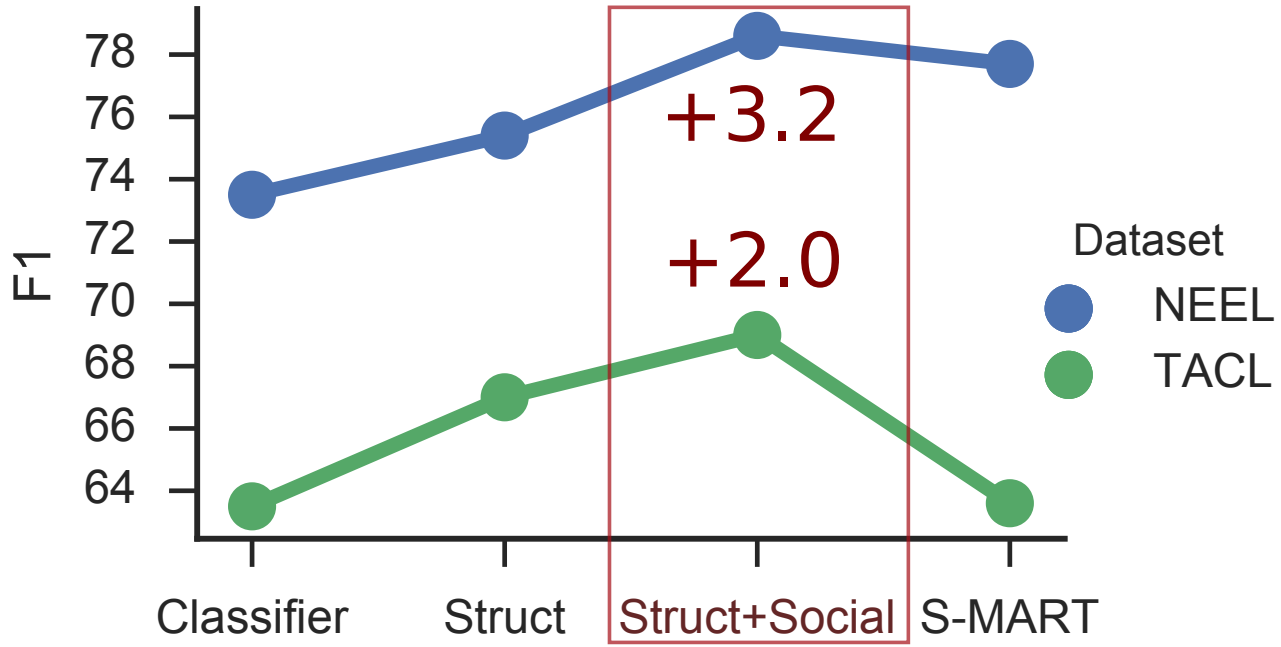
- Optimization: stochastic gradient descent

Inference

- Non-overlapping structure

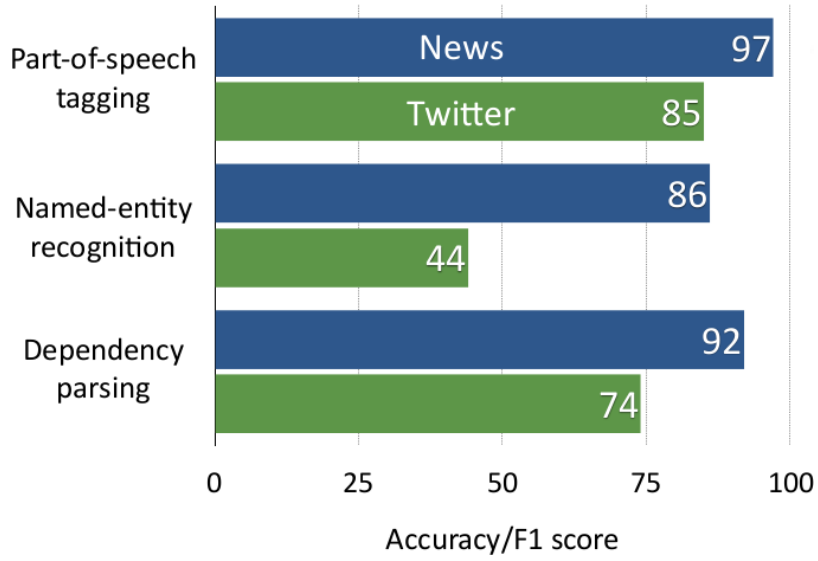


In order to link 'Red Sox' to a real entity, 'Red' and 'Sox' should be linked to Nil.



- Structure prediction improves accuracy.
- Social context yields further improvements.
- S-MART is the prior state-of-the-art (Yang & Chang, 2015).

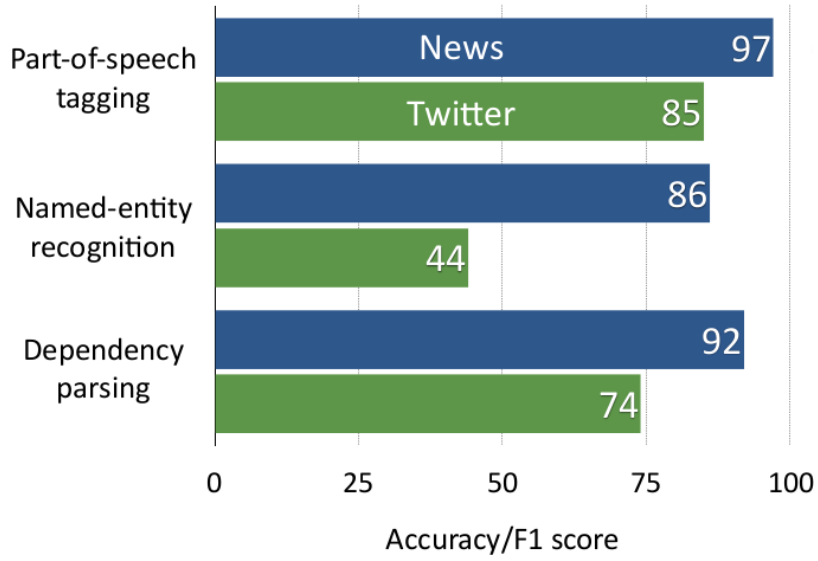
Social media has forced NLP to confront the challenge of missing social context (Eisenstein, 2013):



- 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."

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, $P(y \mid \boldsymbol{x}, a)$, where a is the author.
- Problem: We have labeled examples for only a few authors.

Personalization by ensemble

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

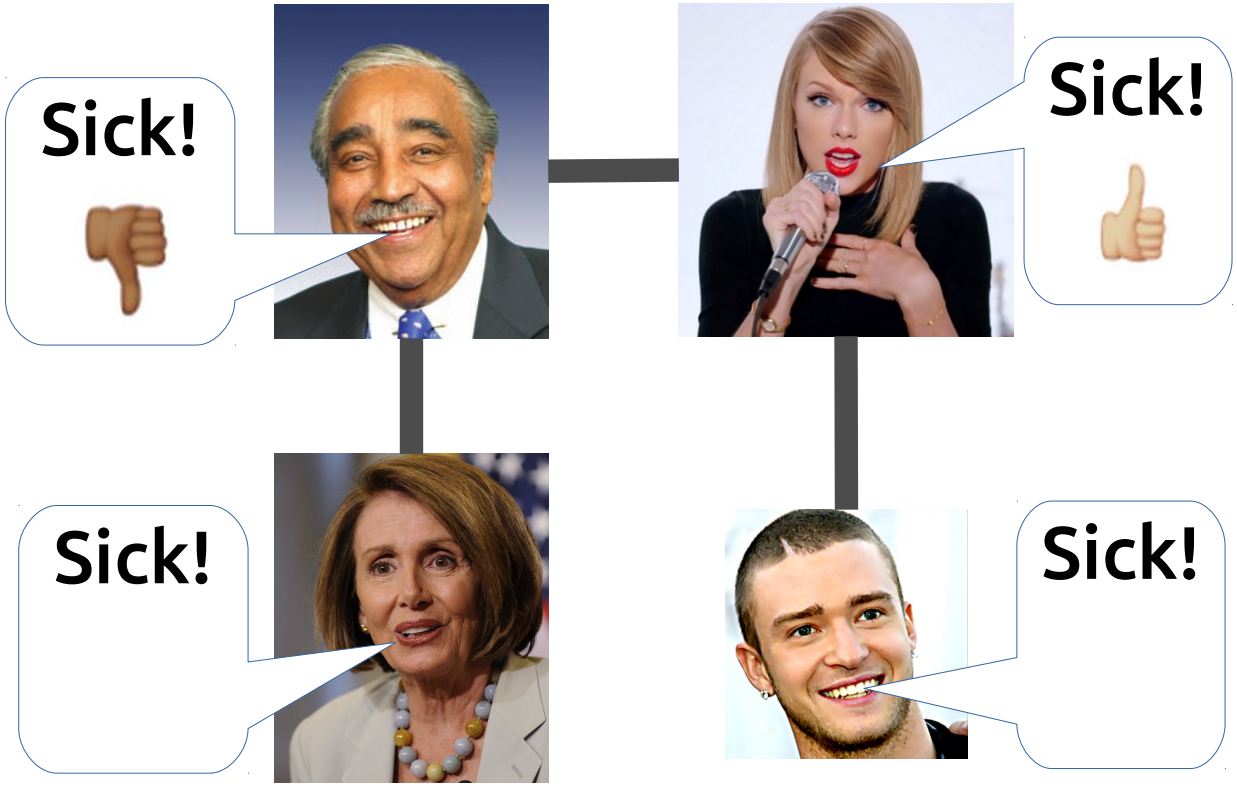
$$P(y \mid \boldsymbol{x}, a) = \sum_k P_k(y \mid \boldsymbol{x}) \pi_a(k)$$

  - $P_k(y \mid \boldsymbol{x})$ 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

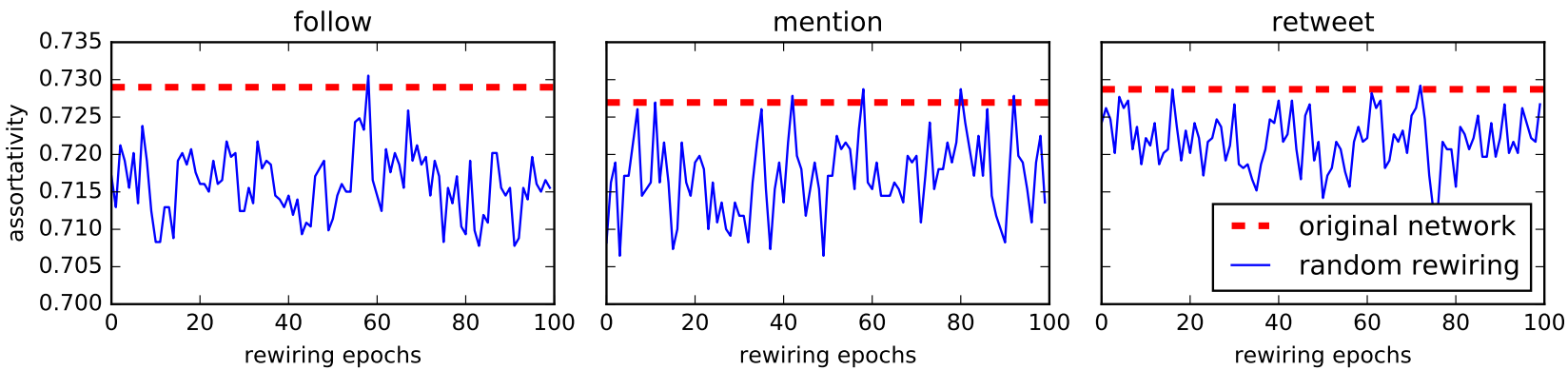
Pilot study: is classifier accuracy **assortative** on the Twitter social network?

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

Evidence for linguistic homophily

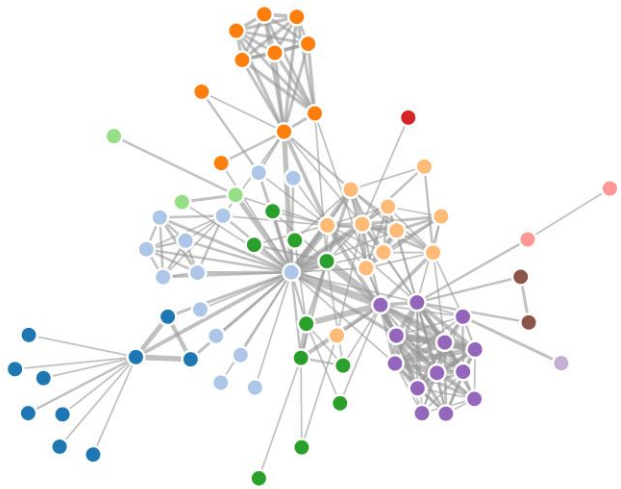
Pilot study: is classifier accuracy **assortative** on the Twitter social network?

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



Network-driven personalization

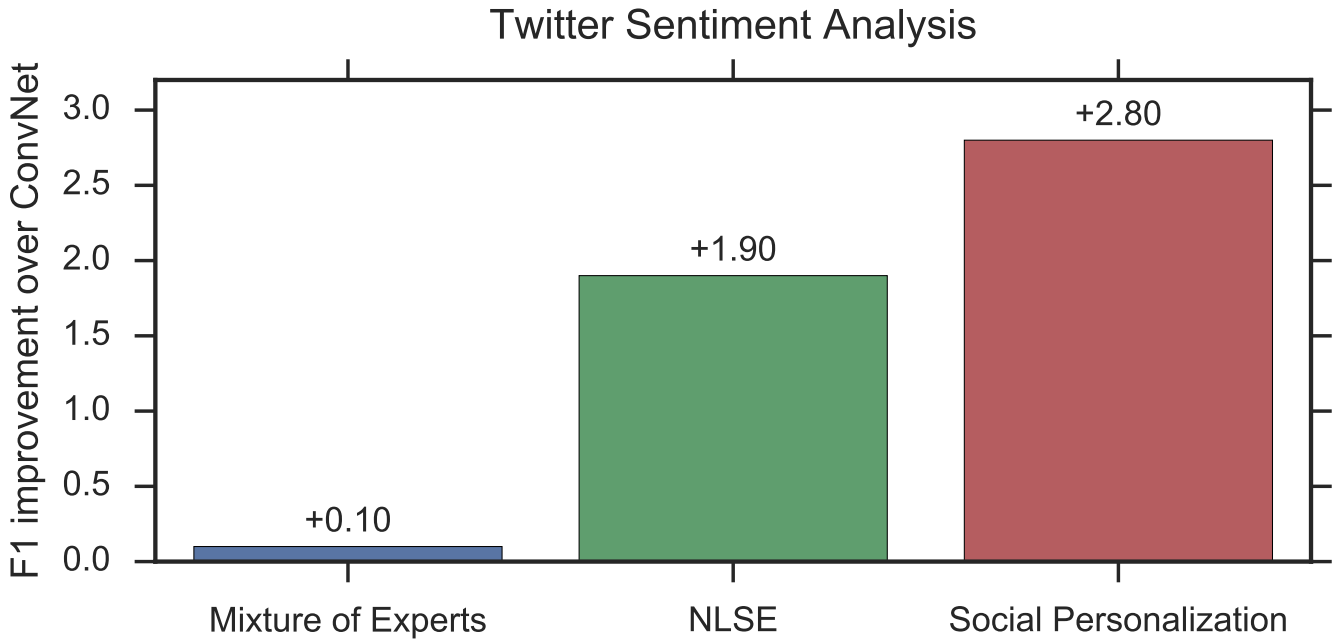
- For each author, estimate a **node embedding** $\boldsymbol{e}_a$ (Tang et al., 2015).
- Nodes who share neighbors get similar embeddings.



$$\boldsymbol{\pi}_a = \text{SoftMax}(f(\boldsymbol{e}_a))$$

$$P(y \mid \boldsymbol{x}, a) = \sum_{k=1}^{K} P_k(y \mid \boldsymbol{x}) \pi_a(k)$$

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

	More positive	More negative
1	banging loss fever broken fucking	dear like god yeah wow
2	chilling cold ill sick suck	satisfy trust wealth strong lmao
3	ass damn piss bitch shit	talent honestly voting win clever
4	insane bawling fever weird cry	lmao super lol haha hahaha
5	ruin silly bad boring dreadful	*lovatics* wish *beliebers arianators kendall*

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.

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.

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 $\rightarrow$ Bloomberg)
  - Mingwei Chang (Google Research)
  - See https://gtnlp.wordpress.com/ for more!
- Funding: National Science Foundation, National Institutes for Health, Georgia Tech

References I

Astudillo, R. F., Amir, S., Lin, W., Silva, M., & Trancoso, I. (2015). Learning word representations from scarce and noisy data with embedding sub-spaces. In *Proceedings of the Association for Computational Linguistics (ACL)*, Beijing.

Eisenstein, J. (2013). What to do about bad language on the internet. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, (pp. 359–369).

Eisenstein, J. (2017). Unsupervised learning for lexicon-based classification. In *Proceedings of the National Conference on Artificial Intelligence (AAAI)*, San Francisco.

Foster, J., Cetinoglu, O., Wagner, J., Le Roux, J., Nivre, J., Hogan, D., & van Genabith, J. (2011). From news to comment: Resources and benchmarks for parsing the language of web 2.0. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, (pp. 893–901)., Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., & Smith, N. A. (2011). Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proceedings of the Association for Computational Linguistics (ACL)*, (pp. 42–47)., Portland, OR.

Pinter, Y., Guthrie, R., & Eisenstein, J. (2017). Mimicking word embeddings using subword rnns. In *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Ritter, A., Clark, S., Mausam, & Etzioni, O. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of EMNLP*.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., & Mei, Q. (2015). Line: Large-scale information network embedding. In *Proceedings of the Conference on World-Wide Web (WWW)*, (pp. 1067–1077).

Yang, Y. & Chang, M.-W. (2015). S-mart: Novel tree-based structured learning algorithms applied to tweet entity linking. In *Proceedings of the Association for Computational Linguistics (ACL)*, (pp. 504–513)., Beijing.

Yang, Y. & Eisenstein, J. (2017). Overcoming language variation in sentiment analysis with social attention. *Transactions of the Association for Computational Linguistics (TACL), in press*.