

Robust NLP Across Dialects and Styles

Jacob Eisenstein
@jacobeisenstein

Georgia Institute of Technology
Facebook AI Research

January 27, 2019

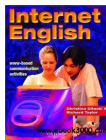
What is a language?

What is a language?

- ▶ A language is a **shared code** of communication.
(lexicon, grammar, phonology, discourse norms, ...)

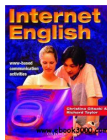
What is a language?

- ▶ A language is a **shared code** of communication.
(lexicon, grammar, phonology, discourse norms, ...)
- ▶ But within a single language, the linguistic phenomena we encounter vary with a number of factors:
 - ▶ when the document was created
 - ▶ the identity of the author/speaker and audience
 - ▶ the communicative setting and genre



What is a language?

- ▶ A language is a **shared code** of communication.
(lexicon, grammar, phonology, discourse norms, ...)
- ▶ But within a single language, the linguistic phenomena we encounter vary with a number of factors:
 - ▶ when the document was created
 - ▶ the identity of the author/speaker and audience
 - ▶ the communicative setting and genre



- ▶ The union of all “Englishes” isn’t understood by anyone!

Consequences for language technology

- ▶ Language tech is built on supervised machine learning. But **all** training sets are incomplete, stale, and biased.
- ▶ Robustness is an endemic problem for NLP:
 - ▶ 20% POS tagging error rate on historical English¹
 - ▶ Poor performance throughout the NLP stack on social media²
 - ▶ Difficulty in generalizing to new domains and genres,³ blocking high-impact applications in health, social sciences, digital humanities, and education.

¹Yang and Eisenstein 2016.

²Ritter et al. 2011; Eisenstein 2013.

³Søgaard 2013.

Solutions?

Solutions?

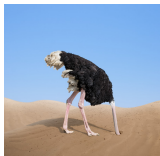


Solutions?



massive, perpetual annotation

Solutions?

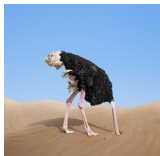


massive, perpetual annotation



domain adaptation

Solutions?



massive, perpetual annotation



domain adaptation



unsupervised pre-training

A manifesto for model-based NLP

Robust generalization from finite, biased data requires stronger inductive biases to identify relationships between observations and labels that are likely to hold in the future.

- ▶ When the test data is “nice,” predictors can get away with confusing causation and correlation.
- ▶ When test data is systematically different from training data, this is much riskier!⁴

⁴Bareinboim and Pearl 2016.

A manifesto for model-based NLP

Robust generalization from finite, biased data requires stronger inductive biases to identify relationships between observations and labels that are likely to hold in the future.

- ▶ When the test data is “nice,” predictors can get away with confusing causation and correlation.
- ▶ When test data is systematically different from training data, this is much riskier!⁴

The necessary inductive biases can be obtained from theory-driven **models** of the text, labels, and covariates.

⁴Bareinboim and Pearl 2016.

Case study: variation in online writing

Social media hosts a particularly diverse array of linguistic phenomena, making it a challenging domain for natural language processing.⁵

- ▶ Many more types of people get to be authors.
- ▶ Relaxed standards of formality.
- ▶ Rich array of communicative contexts.



Tom Brokaw 
@tombrokaw

my tweet portal is whack
i hv been trying to say i am sorry i
offended
and i so appreciate my colleague

5:58 PM · 1/27/19 · [Twitter for iPhone](#)

⁵Ritter et al. 2011; Eisenstein 2013; Baldwin et al. 2013.

Case study: variation in online writing

Social media hosts a particularly diverse array of linguistic phenomena, making it a challenging domain for natural language processing.⁵

- ▶ **Many more types of people get to be authors.**
- ▶ Relaxed standards of formality.
- ▶ Rich array of communicative contexts.



Tom Brokaw 
@tombrokaw

my tweet portal is whack
i hv been trying to say i am sorry i
offended
and i so appreciate my colleague

5:58 PM · 1/27/19 · [Twitter for iPhone](#)


⁵Ritter et al. 2011; Eisenstein 2013; Baldwin et al. 2013.

Personalized classification⁶

- ▶ Goal: personalized conditional likelihood, $P(y \mid \mathbf{x}, a)$, where \mathbf{x} is the text and a is the author.
- ▶ **Personalization ensemble**

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

- ▶ $\pi_a(\cdot)$ are the ensemble weights for author a
- ▶ $P_k(y \mid \mathbf{x})$ is a basis model
(Convolution + MaxPooling + SoftMax)


⁶Yi Yang and Jacob Eisenstein (2017). “Overcoming language variation in sentiment analysis with social attention”. In: *Transactions of the Association for Computational Linguistics (TAGL)* 5. 

Personalized classification⁶

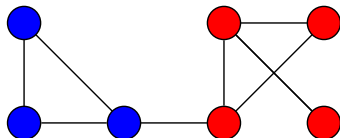
- ▶ Goal: personalized conditional likelihood, $P(y \mid \mathbf{x}, a)$, where \mathbf{x} is the text and a is the author.
- ▶ **Personalization ensemble**

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

- ▶ $\pi_a(\cdot)$ are the ensemble weights for author a
- ▶ $P_k(y \mid \mathbf{x})$ is a basis model
(Convolution + MaxPooling + SoftMax)
- ▶ **Problem**: No labeled examples for most authors.

⁶Yi Yang and Jacob Eisenstein (2017). “Overcoming language variation in sentiment analysis with social attention”. In: *Transactions of the Association for Computational Linguistics (TAGL)* 5. 

Linguistic homophily



- ▶ **Homophily**: socially-connected individuals tend to share traits.⁷
- ▶ Linguistic implication: language differences are correlated with the social network.⁸

⁷McPherson, Smith-Lovin, and Cook 2001.

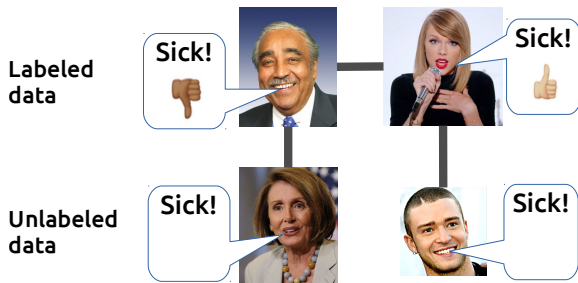
⁸Fagyal et al. 2010; L. Milroy and J. Milroy 1992; Goel et al. 2016.

Linguistic homophily at work

- ▶ I would like to believe he's **sick** rather than just mean and evil.
- ▶ You coulda been getting down to this **sick** beat.

Linguistic homophily at work

- ▶ I would like to believe he's **sick** rather than just mean and evil.
- ▶ You coulda been getting down to this **sick** beat.



Evidence for linguistic homophily

Is predictive accuracy **assortative** on the Twitter social network?

$$\text{assort}(G) = \frac{1}{|G|} \sum_{(i,j) \in G} \underbrace{\delta(y_i = \hat{y}_i) \delta(y_j = \hat{y}_j)}_{\text{correct for both}} + \underbrace{\delta(y_i \neq \hat{y}_i) \delta(y_j \neq \hat{y}_j)}_{\text{incorrect for both}}$$

Evidence for linguistic homophily

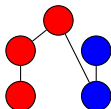
Is predictive accuracy **assortative** on the Twitter social network?

$$\text{assort}(G) = \frac{1}{|G|} \sum_{(i,j) \in G} \underbrace{\delta(y_i = \hat{y}_i) \delta(y_j = \hat{y}_j)}_{\text{correct for both}} + \underbrace{\delta(y_i \neq \hat{y}_i) \delta(y_j \neq \hat{y}_j)}_{\text{incorrect for both}}$$

- ▶ On Twitter, $\text{assort}(G) \approx .73$.
- ▶ Could this happen by chance?

Baseline assortativity

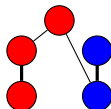
To compare against the null hypothesis of no homophily, we **randomly rewire** the social network and recompute assortativity.



$$\text{assort}(G) = .75$$

Baseline assortativity

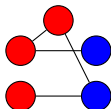
To compare against the null hypothesis of no homophily, we **randomly rewire** the social network and recompute assortativity.



$$\text{assort}(G) = .75$$

Baseline assortativity

To compare against the null hypothesis of no homophily, we **randomly rewire** the social network and recompute assortativity.

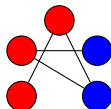


$$\text{assort}(G) = .75$$

$$\text{assort}(\tilde{G}) = (.25, \quad)$$

Baseline assortativity

To compare against the null hypothesis of no homophily, we **randomly rewire** the social network and recompute assortativity.

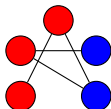


$$\text{assort}(G) = .75$$

$$\text{assort}(\tilde{G}) = (.25, .5, \dots)$$

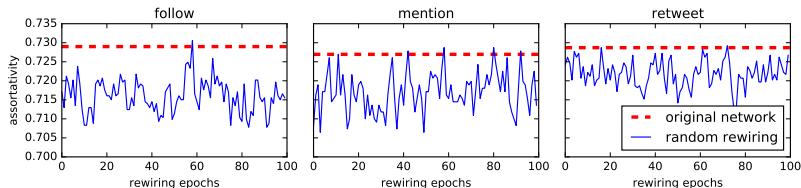
Baseline assortativity

To compare against the null hypothesis of no homophily, we **randomly rewire** the social network and recompute assortativity.



$$\text{assort}(G) = .75$$

$$\text{assort}(\tilde{G}) = (.25, .5, \dots)$$



Social personalization

Personalization ensemble

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

- ▶ $P_k(y \mid \mathbf{x})$ is a basis model
(Convolution + MaxPooling + SoftMax)
- ▶ $\pi_a(\cdot)$ are the ensemble weights for author a
- ▶ **Problem:** No labeled examples for most authors.

Social personalization

Personalization ensemble

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

- ▶ $P_k(y \mid \mathbf{x})$ is a basis model
(Convolution + MaxPooling + SoftMax)
- ▶ $\pi_a(\cdot)$ are the ensemble weights for author a
- ▶ **Problem:** No labeled examples for most authors.
- ▶ **Solution:** impute ensemble weights from the social network structure.

Summarizing network information

For each author, estimate a **node embedding** \mathbf{e}_a , which summarizes the position of the author in the Twitter social network.⁹

$$\max_{\{\mathbf{e}_i\}} \sum_{i,j \in G} \log \sigma(\mathbf{e}_i \cdot \mathbf{e}_j) + \frac{1}{N_s} \sum_{j' \sim P(j)}^{N_s} \log \sigma(-\mathbf{e}_i \cdot \mathbf{e}_{j'}).$$

- ▶ Very similar to word2vec, but for nodes in a network.
- ▶ Neighbors tend to have similar vectors.

⁹Tang et al. 2015.

Network driven personalization

Nodes who share neighbors get similar embeddings.

$$\pi_a = \text{SoftMax}(f(\mathbf{e}_a))$$

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

where $f(\cdot)$ is a multi-layer perceptron.

Network driven personalization

Nodes who share neighbors get similar embeddings.

$$\pi_a = \text{SoftMax}(f(\mathbf{e}_a))$$

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

where $f(\cdot)$ is a multi-layer perceptron.

- ▶ $\{\mathbf{e}_a\}$ is trained offline from the network alone.
- ▶ $f(\cdot)$ and $P_k(y \mid \mathbf{x})$ are trained jointly from labeled data.

Key point: nodes in similar parts of the network get similar classification functions.

The SemEval social network

SemEval 2013-2015 is a standard benchmark for Twitter Sentiment Analysis.¹⁰

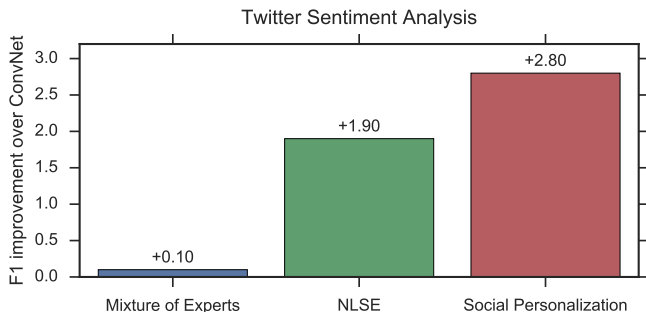
	# Author	# Relations	# Isolates
original	14,087	40,110	3,633
expanded	17,417	1,050,369	689

Table: Follower network statistics

We “densify” the social network by adding individuals who are followed by at least 100 SemEval authors.

¹⁰Nakov et al. 2013; Rosenthal et al. 2015.

Results



Improvements over ConvNet baseline:

- ▶ +2.8% on Twitter Sentiment Analysis
- ▶ +2.7% on Ciao Product Reviews

The three Twitter networks give similar performance, retweet is slightly better.

Variable sentiment words

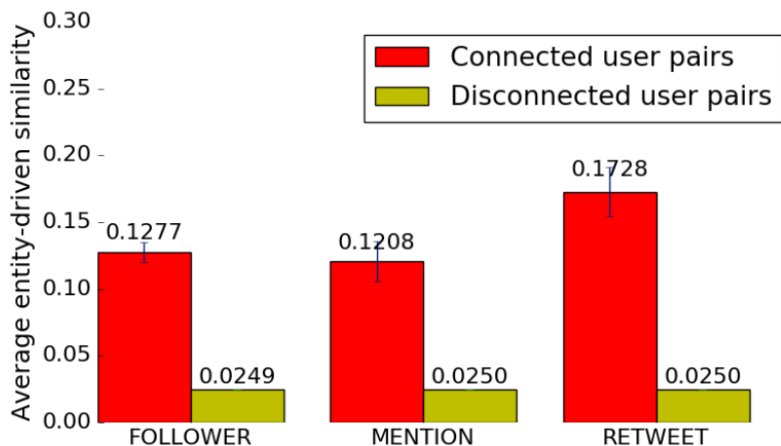
More positive	More negative
1 banging loss fever broken <u>fucking</u>	dear like god yeah wow
2 chilling cold ill sick suck	satisfy trust wealth strong lmao
3 <u>ass</u> <u>damn</u> <u>piss</u> <u>bitch</u> <u>shit</u>	talent honestly voting win clever
4 insane bawling fever weird cry	lmao super lol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

Emergent language styles:

- ▶ Sarcasm
- ▶ Swearing, but in a good way
- ▶ *Haters*

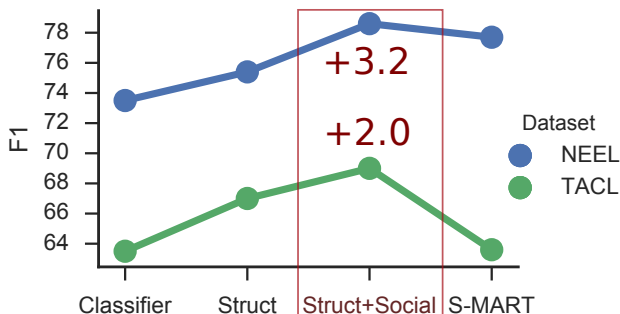
Social personalization for entity linking

Social neighbors tend to talk about the same entities.¹¹



¹¹Yang, Chang, and Eisenstein 2016.

Entity linking results



- ▶ Structure prediction improves accuracy.
- ▶ Social context yields further improvements.
- ▶ S-MART is the prior state-of-the-art¹².

¹²Yang and Chang 2015.

Summary so far

Different people use language differently.

- ▶ Personalized language technology is the holy grail, but we lack labeled data.
- ▶ **Social personalization** links language style to social network structure, improving performance on sentiment analysis and entity linking.

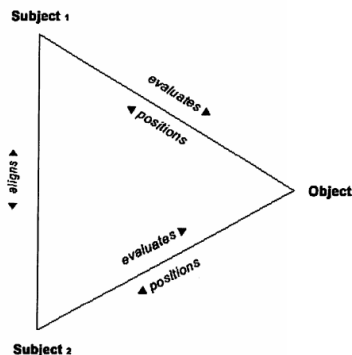
Summary so far

Different people use language differently.

- ▶ Personalized language technology is the holy grail, but we lack labeled data.
- ▶ **Social personalization** links language style to social network structure, improving performance on sentiment analysis and entity linking.

What about when a single person uses language differently across contexts?

Interactional stancetaking¹³



- ▶ **Affect**: attitude towards stance object
- ▶ **Alignment**: attitude towards audience
- ▶ **Investment**: attitude towards talk

¹³Scott F. Kiesling et al. (2018). "Interactional Stancetaking in Online Forums". In: *Computational Linguistics* 44 (4).

A stancetaking corpus

Pilot corpus: annotations of the three stance dimensions on a 1-5 scale on 70 Reddit threads (~ 1400 turns)

ID-rep	Content	Stance focus	User	Karma	Affect	Investment	Alignment
008-02	People that are truly bad servers don't last very long at restaurants anyway. It is one of the jobs you actually need to be good at, you cant fake it like an engineer.	faking it like engineers	A ₄	-2	3	3	3
009	You can't fake engineering	faking engineering	A ₅	2	3	4	1
010	Lol... oh yes you can. You have to get the degree but that doesn't mean you are any use to anyone at your job.	faking engineering	A ₄	2	3	3	1

Stance keywords

AFFECT		INVESTMENT		ALIGNMENT	
HIGH	LOW	HIGH	LOW	HIGH	LOW
thank ! sing noise stop	please worse everyone nothing entire	! tell hope better never	little limit ink maybe may	thank limit other ! absolutely	evidence wrong able not opinion

Stance keywords

AFFECT		INVESTMENT		ALIGNMENT	
HIGH	LOW	HIGH	LOW	HIGH	LOW
thank ! sing noise stop	please worse everyone nothing entire	! tell hope better never	little limit ink maybe may	thank limit other ! absolutely	evidence wrong able not opinion

Next steps:

- ▶ connections to constructs like politeness and formality;
- ▶ a stancetaking dialogue system;
- ▶ translations that convey the appropriate interactional stance.

Stylistic variation in translation¹⁴

A: This sat work??

B: I kinda had plans
already. Wassup
with a weekday

A: I gotta beachouse
4th of July if y'all
wanna come up

B: Ehhh I have off
the 27th so maybe
the 26th

¹⁴Song et al. 2018.

Stylistic variation in translation¹⁴

A: This sat work??	Ce travail assis??	This job sitting??
B: I kinda had plans already. Wassup with a weekday	J'ai un peu deja des plans. Wassup avec un jour de semaine	I already have some plans. Wassup with a weekday
A: I gotta beachouse 4th of July if y'all wanna come up	Je dois beachouse le 4 juillet si vous voulez monter	I have beachouse on July 4th if you want to ride
B: Ehhh I have off the 27th so maybe the 26th	Ehhh j'ai le 27 alors peut-être le 26	Ehhh I have the 27th so maybe the 26th

¹⁴Song et al. 2018.

Giving MT a social life

- ▶ Very limited public data for measuring translation of SMS-style text.¹⁵
 - ▶ This summer: a team at the Jelinek summer workshop will (hopefully) address this problem!

¹⁵Michel and Neubig 2018.

¹⁶Belinkov and Bisk 2018.

Giving MT a social life

- ▶ Very limited public data for measuring translation of SMS-style text.¹⁵
 - ▶ This summer: a team at the Jelinek summer workshop will (hopefully) address this problem!
- ▶ How to handle character-level variation in MT?¹⁶ (e.g., eh hh, wassup)
 - ▶ Pinter, Guthrie, and E (2017): “mimicking” word embeddings with sub-word RNNs
 - ▶ This ACL (?): making character-level encoders more robust to natural noise, such as spelling errors.

¹⁵Michel and Neubig 2018.

¹⁶Belinkov and Bisk 2018.

Conclusion: embrace diversity!

- ▶ We need to think about language as **fluid and diverse**, not **fixed and monolithic**.
- ▶ Fortunately, variation and change are not random noise.
- ▶ By **modeling** the underlying sociolinguistic drivers of language variation and change, we can build a new generation of robust language technology.

Acknowledgments

- ▶ **Students:** Yangfeng Ji, Yi Yang, *Umashanthi Pavalanathan*, Sandeep Soni, Ian Stewart, Yuval Pinter, Sarah Wiegreffe, *Murali Basulu*, *Taha Merghani*, *Xiaochuang Han*
- ▶ **Collaborators:** Stergos Afantenos, *Ming-Wei Chang*, Munmun De Choudhury, Eric Gilbert, *Scott Kiesling*, Lauren F. Klein, Dong Nguyen, Brendan O'Connor, Noah A. Smith, Manfred Stede, Eric P. Xing
- ▶ **Sponsors:** NSF, AFOSR, NIH, DTRA, NEH, Google

References I



Astudillo, Ramon F et al. (2015). “Learning Word Representations from Scarce and Noisy Data with Embedding Sub-spaces”. In: *Proceedings of the Association for Computational Linguistics (ACL)*.



Baldwin, Timothy et al. (2013). “How noisy social media text, how diffrent social media sources”. In: *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pp. 356–364.



Bareinboim, Elias and Judea Pearl (2016). “Causal inference and the data-fusion problem”. In: *Proceedings of the National Academy of Sciences* 113.27, pp. 7345–7352.



Belinkov, Yonatan and Yonatan Bisk (2018). “Synthetic and natural noise both break neural machine translation”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*.



Eisenstein, Jacob (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.

References II

- 
- Fagyal, Zsuzsanna et al. (2010). "Centers and peripheries: Network roles in language change". In: *Lingua* 120.8, pp. 2061–2079.
- 
- Goel, Rahul et al. (Nov. 2016). "The Social Dynamics of Language Change in Online Networks". In: *The International Conference on Social Informatics (SocInfo)*.
- 
- Kiesling, Scott F. et al. (2018). "Interactional Stancetaking in Online Forums". In: *Computational Linguistics* 44 (4).
- 
- McPherson, Miller, Lynn Smith-Lovin, and James M Cook (2001). "Birds of a feather: Homophily in social networks". In: *Annual review of sociology* 27, pp. 415–444.
- 
- Michel, Paul and Graham Neubig (2018). "MTNT: A Testbed for Machine Translation of Noisy Text". In: *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 543–553.
- 
- Milroy, Lesley and James Milroy (1992). "Social network and social class: Toward an integrated sociolinguistic model". In: *Language in society* 21.01, pp. 1–26.

References III



Nakov, Preslav et al. (2013). “Semeval-2013 task 2: Sentiment analysis in twitter”. In: *Proceedings of the 7th International Workshop on Semantic Evaluation, SemEval*.



Pinter, Yuval, Robert Guthrie, and Jacob Eisenstein (2017). “Mimicking Word Embeddings using Subword RNNs”. In: *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.



Ritter, Alan et al. (2011). “Named entity recognition in tweets: an experimental study”. In: *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.



Rosenthal, Sara et al. (2015). “Semeval-2015 task 10: Sentiment analysis in twitter”. In: *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval*.





Søgaard, Anders (2013). “Semi-Supervised Learning and Domain Adaptation in Natural Language Processing”. In: *Synthesis Lectures on Human Language Technologies* 6.2, pp. 1–103.



Song, Zhiyi et al. (2018). *BOLT English SMS/chat LDC2018T19*. Philadelphia.

References IV

-  Tang, Jian et al. (2015). “Line: Large-scale information network embedding”. In: *Proceedings of the Conference on World-Wide Web (WWW)*, pp. 1067–1077.
-  Yang, Yi and Ming-Wei Chang (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.
-  Yang, Yi, Ming-Wei Chang, and Jacob Eisenstein (2016). “Toward Socially-Infused Information Extraction: Embedding Authors, Mentions, and Entities”. In: *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.
-  Yang, Yi and Jacob Eisenstein (2016). “Part-of-Speech Tagging for Historical English”. In: *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
-  – (2017). “Overcoming language variation in sentiment analysis with social attention”. In: *Transactions of the Association for Computational Linguistics (TACL)* 5.