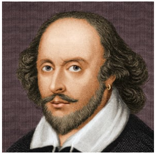
Making Fetch Happen

Language Change in Social and Linguistic Context

Jacob Eisenstein

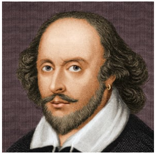
Change as a constant

Full fathom five thy father lies; Of his bones are coral made.

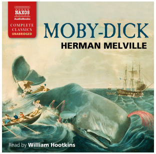


Change as a constant

Full fathom five thy father lies; Of his bones are coral made.

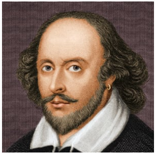


Aye, aye! it was that accursed white whale that razed me; made a poor pegging lubber of me for ever and a day!

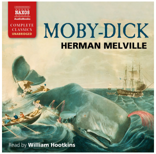


Change as a constant

Full fathom five thy father lies; Of his bones are coral made.



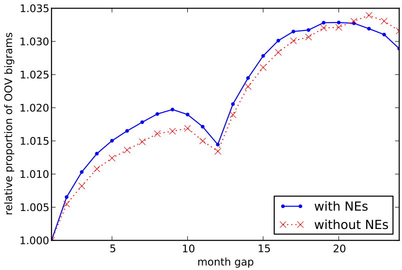
Aye, aye! it was that accursed white whale that razed me; made a poor pegging lubber of me for ever and a day!



Now if you'll excuse me, I'm going to go on an overnight drunk, and in 10 days I'm going to set out to find the shark that ate my friend and destroy it.

Short-term change

Change happens by month, not just by decade![1]



[1] Eisenstein 2013.

Language change and NLP

Natural language processing hasn't taken language change very seriously.

- Existing corpora are usually drawn from narrow periods of time, mostly since 1990s.
- Performance on historical texts is poor:
  25% error rate on POS tagging for early modern English.[2][3]
- Poor performance on contemporary social media text is also partly due to inability to adapt to language change.

---

[2] Yang and Eisenstein 2016.

[3] Ask me about recent results adapting BERT for Early Modern English! (Han and Eisenstein 2019)

Language change and sociolinguistics

Weinreich, Labov, and Herzog (1968) present five problems:

- **Constraints**: what changes are possible?
- **Transition**: how does a change propagate in a community of speakers?
- **Embedding**: what implications does a change have for the larger linguistic system?
- **Evaluation**: what is the social meaning of a particular change?
- **Actuation**: why this change, and why now?

Language change and sociolinguistics

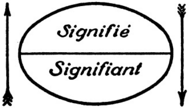
Weinreich, Labov, and Herzog (1968) present five problems:

- **Constraints**: what changes are possible?
- **Transition**: how does a change propagate in a community of speakers?
- **Embedding**: what implications does a change have for the larger linguistic system?
- **Evaluation**: what is the social meaning of a particular change?
- **Actuation**: why this change, and why now?

What can diachronic data tell us about social structures? About the organization of the linguistic system?

The Dynamic Lexicon[4]

Lexical innovation can happen on the level of new wordforms (signs) and new meanings (signifieds).

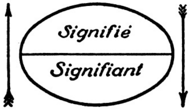


- Changes in a corpus may be driven by new real-world events and entities (e.g., email, #viadoom).
- Linguistic "fashions" involve new signs for existing meanings (lol).
- In other cases, existing signs get repurposed to new meanings (hot, fetch).

[4] Pierrehumbert 2010.

The Dynamic Lexicon[4]

Lexical innovation can happen on the level of new wordforms (signs) and new meanings (signifieds).

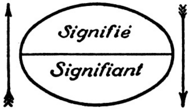


- Changes in a corpus may be driven by new real-world events and entities (e.g., email, #viadoom).
- **Linguistic "fashions" involve new signs for existing meanings (lol).**
- In other cases, existing signs get repurposed to new meanings (hot, fetch).

[4] Pierrehumbert 2010.

The Dynamic Lexicon[4]

Lexical innovation can happen on the level of new wordforms (signs) and new meanings (signifieds).



- Changes in a corpus may be driven by new real-world events and entities (e.g., email, #viadoom).
- Linguistic "fashions" involve new signs for existing meanings (lol).
- **In other cases, existing signs get repurposed to new meanings (hot, fetch).**

[4] Pierrehumbert 2010.

Making fetch happen

The influence of social and linguistic context on nonstandard word growth and decline[5]

> Stop trying to make "fetch" happen! It's not going to happen!
>
> *Regina George, Mean Girls (2005)*

[5] Ian Stewart and Jacob Eisenstein (2018). "Making "fetch" happen: The influence of social and linguistic context on the success of lexical innovations". In: *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

Background

What factors predict whether an innovative slang term will succeed or fail?

- Prior work has focused largely on **social factors**: who are the early adopters, how is their social network organized, and how influential are they?[6]
- This work considers **linguistic factors**: how does the innovation fit into the existing linguistic system?

[6] Altmann, Pierrehumbert, and Motter 2011; Garley and Hockenmaier 2012.

Social dissemination

Altmann, Pierrehumbert, and Motter (2011): successful innovations disseminate widely across social contexts.

- For example, it is better to have three adopters in three cities than in one city.

[7] Garley and Hockenmaier 2012.

[8] Altmann, Pierrehumbert, and Motter 2011.

Social dissemination

Altmann, Pierrehumbert, and Motter (2011): successful innovations disseminate widely across social contexts.

- For example, it is better to have three adopters in three cities than in one city.

Quantifying dissemination:

$$D = \log \frac{\text{count-of-contexts}}{E[\text{count-of-contexts} \mid \text{total-counts}]} \tag{1}$$

- one context = one user[7]
- one context = one newsgroup[8]

[7] Garley and Hockenmaier 2012.

[8] Altmann, Pierrehumbert, and Motter 2011.

Linguistic dissemination

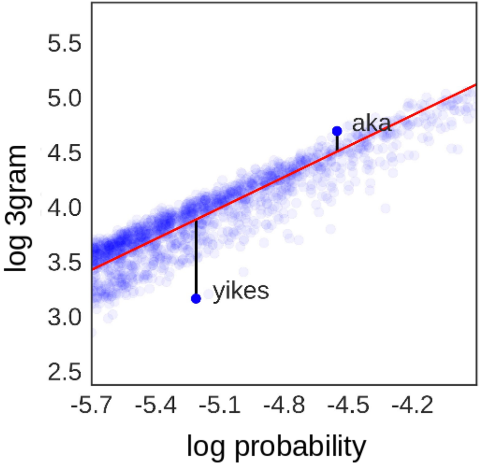
- This and other prior work treats language no differently from hashtags[9] or hyperlinks.[10] But language is different, because innovations must interact with the rest of the linguistic system.
- Our hypothesis is that linguistically versatile innovations tend to succeed. We define **linguistic dissemination**: one context = one trigram.

$$D^{(\ell)} = \log \frac{\text{count-of-trigrams}}{E[\text{count-of-trigrams} \mid \text{total-counts}]} \tag{2}$$

[9] Romero, Meeder, and Kleinberg 2011.

[10] Bakshy et al. 2012.

Linguistic dissemination



Data

- 1.6B public Reddit posts and comments from 2013-2016
  - Filtered known bots and spammers[11]
  - English-language subreddits only
- Vocabulary methodology: automatically search, manually filter.[12]
  1. Automatically identify words with consistent growth for at least part of the data.
  2. Manually filter out proper nouns and standard English ($\kappa = .79$).

[12] Tan and Lee 2015.

[12] Eisenstein, O'Connor, et al. 2014.

Examples

	Word	Gloss	Formation type	
growth	idk	I don't know	acronym	N=1120
	shitpost	low-quality post	compound	
	tho	though	clipping	
decline	eyebleach	pleasing image(s)	compound	N=530
	trashy	undesirable	derivation	
	wot	what	respelling	

Analyses

1. Does (linguistic/social) dissemination **cause** word frequency to increase?
2. Can dissemination help to **predict**
   - which words will increase in frequency?
   - how long each innovation will survive?

Causal analysis

Potential outcomes perspective: "if this individual had/hadn't been treated, what would have been the outcome?" In this case:

- **Treatment**: amount of dissemination;
- **Outcome**: whether word increases in frequency after 12 months;
- **Covariates**: everything else we know about each word.

Propensity score matching is a well-known approach to this problem,[13] but extra care is required when the treatment is continuous.

[13] Rosenbaum and Rubin 1983.

Average dose-response function[14]

1. Fit a model of the treatment from the covariates,

$$Z_i \mid X_i \sim N(\beta \cdot x_i, \sigma_Z^2). \tag{3}$$

   The generalized propensity score R_i is the conditional likelihood $P(z_i \mid x_i)$.

2. Regress the outcome against the treatment and the generalized propensity score,

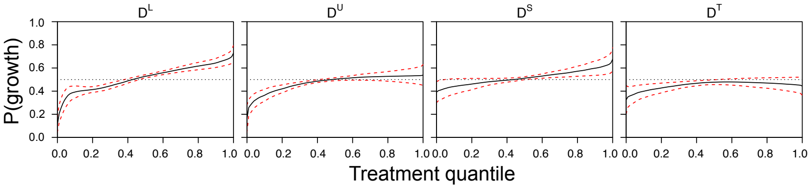
$$\hat{Y}_i = \sigma(\hat{\alpha}_0 + \hat{\alpha}_1 Z_i + \hat{\alpha}_2 R_i). \tag{4}$$

3. At each treatment quantile, s_z, compute the average predicted outcome for each instance,

$$\hat{\mu}(s_z) = \frac{1}{|s_z|} \sum_{i : z_i \in s_z} \hat{Y}_i. \tag{5}$$

[14] Hirano and Imbens 2004.

Average dose-response results



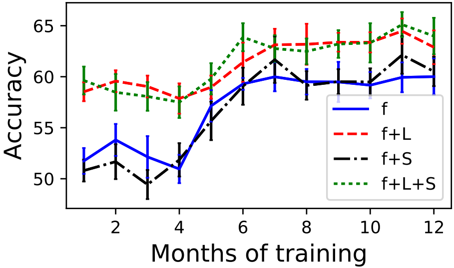
- Linguistic dissemination (D^L) steadily increases the probability that an innovation will be adopted (left).
- Of the three social dissemination indicators, only subreddit dissemination (D^S) makes a significant impact on adoption.

Predicting word success

Given t months of training data, can we predict whether a word will continue to increase in frequency?

Predicting word success

Given t months of training data, can we predict whether a word will continue to increase in frequency?



- f: frequency
- L: linguistic dissemination
- S: social dissemination

Predicting word survival

Can we predict when innovations will start to lose popularity?

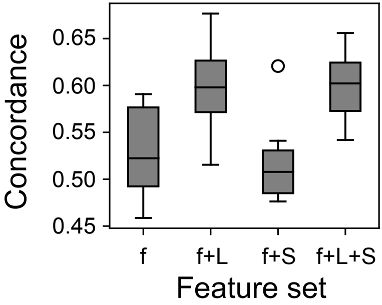
- Cox proportional hazards model,

$$\lambda_i(t) = \lambda_0(t) \exp(\beta \cdot \mathbf{x}_i), \tag{6}$$

  where
  - $\lambda_i(t)$ is the hazard of "death" at time t;
  - $\mathbf{x}_i$ is a vector of predictors;
  - β is a vector of weights.
- Must adjust for right-censored data, since not all innovations decline during our sample.

Predicting word survival

- Of all the dissemination statistics, only linguistic dissemination is a statistically significant predictor of survival.
- Including linguistic dissemination significantly increases predictive accuracy (as measured by concordance).



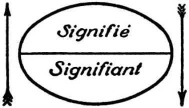
Summary of this part

- Successful innovations disseminate into a diverse set of phrases, rather than a few popular fixed expressions.
- After accounting for linguistic dissemination, social dissemination is a weak predictor at best.
- Linguistic innovations can help to measure social phenomena, but they are different from other types of innovations, like hashtags, hyperlinks, and formatting conventions.[15]

[15] Rotabi, Danescu-Niculescu-Mizil, and Kleinberg 2017.

The Dynamic Lexicon[4]

Lexical innovation can happen on the level of new wordforms (signs) and new meanings (signifieds).

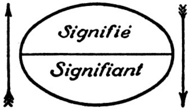


- Changes in a corpus may be driven by new real-world events and entities (e.g., email, #viadoom).
- **Linguistic "fashions" involve new signs for existing meanings (lol).**
- In other cases, existing signs get repurposed to new meanings (hot, fetch).

[4] Pierrehumbert 2010.

The Dynamic Lexicon[4]

Lexical innovation can happen on the level of new wordforms (signs) and new meanings (signifieds).



- Changes in a corpus may be driven by new real-world events and entities (e.g., email, #viadoom).
- Linguistic "fashions" involve new signs for existing meanings (lol).
- **In other cases, existing signs get repurposed to new meanings (hot, fetch).**

[4] Pierrehumbert 2010.

Quantifying Semantic Progressiveness of Documents[16]

[16] Sandeep Soni, Kristina Lerman, and Jacob Eisenstein (2019). "Quantifying Semantic Progressiveness of Documents". In: *submitted to ACL*.

Follow the leader?

- Languages change by assigning new meanings to existing signs.[17]
- Recent work on **diachronic word embeddings** can capture such changes.[18]
- Can we identify **documents** that lead semantic changes? Are these documents especially influential?

[17] Traugott and Dasher 2001.

[18] Kulkarni et al. 2015; Hamilton, Leskovec, and Jurafsky 2016b; Rosenfeld and Erk 2018.

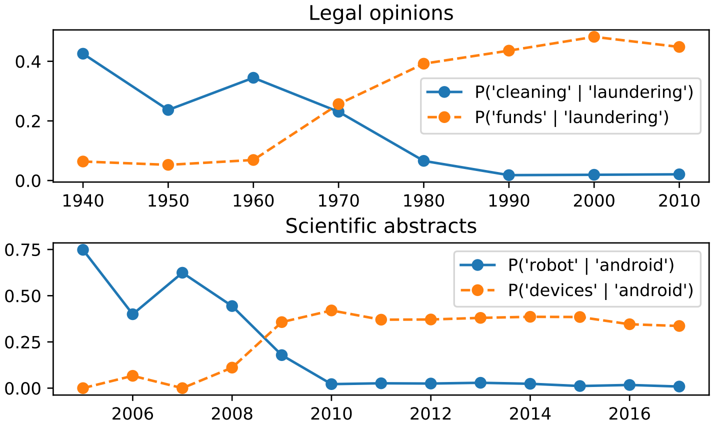
Diachronic word embeddings

- Word embeddings are vector representations of word meaning.[19]
- Which words changed their meanings?[20]
  1. Let $\mathcal{N}_w^{(t)}$ be the near-neighbors of word w at time t.
  2. A word undergoes semantic change when $|\mathcal{N}_w^{(t)} \cap \mathcal{N}_w^{(t+1)}|$ is small.

[19] Mikolov et al. 2013.

[20] Hamilton, Leskovec, and Jurafsky 2016a.

Examples


Legal opinions
0.0
0.2
0.4
1940
1950
1960
1970
1980
1990
2000
2010
P('cleaning' | 'laundering')
P('funds' | 'laundering')
Scientific abstracts
0.00
0.25
0.50
0.75
2006
2008
2010
2012
2014
2016
P('robot' | 'android')
P('devices' | 'android')

Identifying progressive usages

- Is a given usage more likely to be the "old" or "new" meaning?
- The skipgram word embedding model computes the probability of the context around each word,

$$\log P(w_{i+k} \mid w_i) = \mathbf{v}_{w_{i+k}} \cdot \mathbf{u}_{w_i} - \log \sum_{w'} \exp \mathbf{v}_{w'} \cdot \mathbf{u}_{w_i}. \quad (7)$$

- The "progressiveness" of a usage is the log-odds ratio,

$$r_t \triangleq \sum_k \log \frac{P^{(\text{new})}(w_{i+k} \mid w_i)}{P^{(\text{old})}(w_{i+k} \mid w_i)}. \quad (8)$$

The progressiveness of a document (with respect to a single word) is the sum of this statistic.

Examples

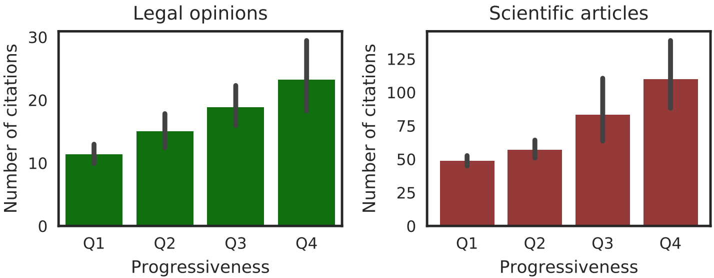
Corpus	Innovation	Leading document
Legal	laundering	United States v. Talmadge G. Rauhoff (7th Cir. 1975)
	asylum	Bertrand v. Sava (S.D.N.Y. 1982)
	fertilization	Planned Parenthood vs Casey (505 U.S. 833)
Science	ux	Hassenzahl and Tractinsky (2006)
	surf	Bay et al (2008)
	android	Shabtai et al (2010)

Examples

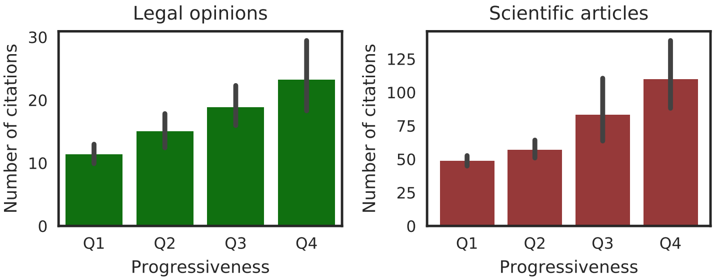
Corpus	Innovation	Leading document
Legal	laundering	United States v. Talmadge G. Rauhoff (7th Cir. 1975)
	asylum	Bertrand v. Sava (S.D.N.Y. 1982)
	fertilization	Planned Parenthood vs Casey (505 U.S. 833)
Science	ux	Hassenzahl and Tractinsky (2006)
	surf	Bay et al (2008)
	android	Shabtai et al (2010)

- ... two-week gestational increments from **fertilization** to full term ...
- ... $15,000 as part of the **'laundering'** process ...
- ... first step in the successful **laundering** of the funds...

Do semantic leaders get more citations?



Do semantic leaders get more citations?



These differences are still significant in a multivariate regression controlling for age, length, out-citations, and number of unique terms.

You can't stay here

The Effectiveness of Reddit's 2015 Ban Through the Lens of Hate Speech[21]

[21] Eshwar Chandrasekharan et al. (2018). "You Can't Stay Here: The Effectiveness of Reddit's 2015 Ban Through the Lens of Hate Speech". In: *Proceedings of Computer-Supported Cooperative Work (CSCW)*.

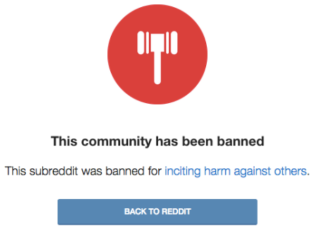
Hate speech on Reddit

What happens when forums for hate speech are shut down?

- Do participants export hate speech elsewhere?
- Or does the elimination of the "echo chamber" reduce hate speech overall?

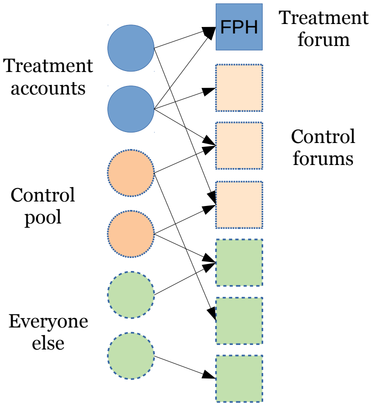
A natural experiment

- In 2015, Reddit closed several forums for violations of its anti-harassment policy.
- This enables a **natural experiment** on the effectiveness of this intervention.



Causal inference design

- **Treatment group**: user accounts that post in the forums that were banned
- **Control forums**: other forums where the treatment group posts
- **Control pool**: other accounts who post in the control forums
- **Control group**: user accounts selected by Mahalanobis Distance Matching in the control pool



Measuring hate speech

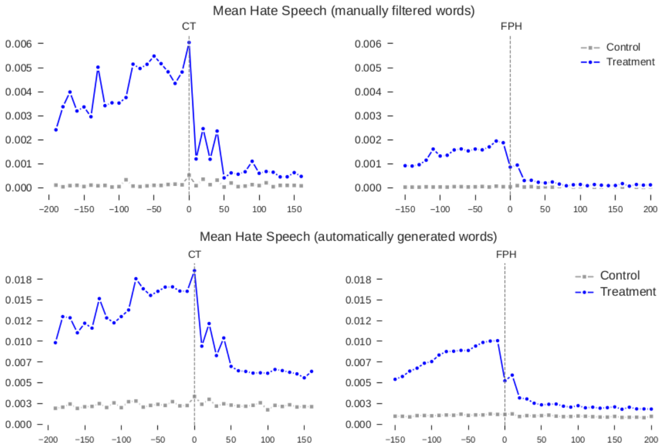
1. Identify words that are unusually frequent in each forum, using SAGE.[22].

2. Examine the top 100, manually remove words that are not intrinsically linked to hate speech (EU Court of Human Rights definition)
   - the forum itself: fph, ct
   - the act of posting offensive content: shitposting, shitlord
   - words often used in non-hate speech contexts: IQ, welfare, cellulite

   High interrater agreement, $\kappa \approx .88$

[22] Eisenstein, Ahmed, and Xing 2011.

Causal effect on hate speech



Aftermath

47.0k

Reddit's bans of r/coontown and r/fatpeoplehate worked--many accounts of frequent posters on those subs were abandoned, and those who stayed reduced their use of hate speech (comp.social.gatech.edu)

5 months ago by asbruckman

Professor | Interactive Computing x2

6649 comments share save hide report

Aftermath

[–] **Hey-Grandan2** 349 points 5 days ago

What excacly qualifies for hate speech?

permalink embed save report give gold reply

[–] **eegilbert** Author of Article 652 points 5 days ago

One of the authors here. There was an unsupervised computational process used, documented on pages 6 and 7, and then a supervised human annotation step. Both lexicons are used throughout the rest of work.

permalink embed save parent report give gold reply

[+] *Comment removed 5 days ago* (58 children)*

[–] **Laminar_flo** 92 points 5 days ago

Ok, adding to that, how did you ensure that the manual filtering process was ideological neutral and not just a reflection of the political sensitivities of the person filtering?

permalink save parent report give gold reply

[–] **qwenjwenfljnanq** 11 points 5 days ago

But then how did you differentiate between hate speech and people talking *about* hate speech?

permalink save parent report give gold reply

[–] **Mode1961** -14 points 5 days ago

> number of words that indicate hate speech

Who choose those words.

permalink save parent report give gold reply

Aftermath

U.S.

Reddit Bans Nazi Groups and Others in Crackdown on Violent Content

By CHRISTINE HAUSER OCT. 26, 2017

Steve Huffman, a co-founder and chief executive of Reddit, in 2016. The company has started to implement a new policy to remove content that glorifies and incites violence from its site. David Paul Morris/Bloomberg

RELATED COVERAGE

ON TECHNOLOGY
How Hate Groups Forced Online Platforms to Reveal Their True Nature AUG. 21, 2017

Opinion | Op-Ed Contributor
My Time Undercover With the Alt-Right
SEPT. 27, 2017

THE SHIFT
This Was the Alt-Right's Favorite Chat App. Then Came Charlottesville. AUG. 15, 2017

THE SHIFT
Reddit Limits Noxious Content by Giving Trolls Fewer Places to Gather SEPT. 25, 2017

(Why) did it work for Reddit?

- Reddit's federated structure delegates norm enforcement to moderators.

  It would be hard for Facebook and Twitter to target hate speech *communities* in the same way

- Some users went to alternative sites like Voat.

  Still a win for Reddit?

- Our algorithms detect only specific subsets of hate speech.

  Did hate speech shift to a form that is harder to detect?

Unsupervised Domain Adaptation of Contextualized Embeddings

A Case Study in Early Modern English[23]

[23] Xiaochuang Han and Jacob Eisenstein (2019). "Unsupervised Domain Adaptation of Contextualized Embeddings: A Case Study in Early Modern English". In: *arXiv preprint arXiv:1904.02817*.

Tagging Early Modern English

- *Early Modern English* (EME): 15-17c, contemporaneous with Shakespeare.
- Syntactic annotations available from the Penn-Helsinki Corpus of Historical English.[24]
- Accurate syntactic analysis of historical texts would facilitate research in the digital humanities and historical linguistics.[25]

[24] Kroch, Santorini, and Diertani 2004.

[25] Degaetano-Ortlieb 2018; Muralidharan and Hearst 2013; Vuillemot et al. 2009.

Challenges for NLP

Spelling is the most salient difference from contemporary modern English:

(1) If this marsch waulle were not kept, and the canales of eche partes of Sowey river kept from abundance of wedes, al the plaine marsch ground at sodaine raynes wold be overflowen, and the profite of the meade lost.

Challenges for NLP

Spelling is the most salient difference from contemporary modern English:

(4) If this marsch waulle were not kept, and the canales of eche partes of Sowey river kept from abundance of wedes, al the plaine marsch ground at sodaine raynes wold be overflowen, and the profite of the meade lost.

Other differences include thou and ye pronouns, -th suffix, and inconsistent capitalization.

(5) And that those **Writs** which shall be awarded and directed for returning of **Juryes** . . .

(6) . . . shall not then have **Twenty** pounds or **Eight** pounds respectively . . .

Tagging from contextualized word embeddings

- ELMO and BERT are *contextualized embeddings*: vector representations of each word's role in context, based on pretraining from large-scale unlabeled data.[26]
- For many problems, state-of-the-art results can be achieved by applying a classification layer directly to the contextualized embeddings:

$$\vec{z}_{1:T} = \textsc{Embed}(\vec{x}_{1:T})$$
$$P(y_t \mid \vec{x}_{1:T}) = \textsc{SoftMax}(\Theta \vec{z}_t).$$

[26] Peters et al. 2018; Devlin et al. 2018.

Learning settings for BERT-based tagging

$$\begin{aligned}\vec{z}_{1:T} =& \textsc{Embed}(\vec{x}_{1:T})\\ P(y_t \mid \vec{x}_{1:T}) =& \textsc{SoftMax}(\Theta \vec{z}_t).\end{aligned}$$

- **Pretraining**: the embedding function is learned from a language modeling objective on unlabeled data.
- **Direct transfer**: the embedding function is fixed, and only Θ is learned from labeled data.
- **Fine-tuning**: both the embedding function and Θ are updated during training.

Learning settings for BERT-based tagging

$$\vec{z}_{1:T} = \textsc{Embed}(\vec{x}_{1:T})$$
$$P(y_t \mid \vec{x}_{1:T}) = \textsc{SoftMax}(\Theta \vec{z}_t).$$

- **Pretraining**: the embedding function is learned from a language modeling objective on unlabeled data.
- **Direct transfer**: the embedding function is fixed, and only Θ is learned from labeled data.
- **Fine-tuning**: both the embedding function and Θ are updated during training.

Learning settings for BERT-based tagging

$$\begin{aligned}\vec{z}_{1:T} =& \textsc{Embed}(\vec{x}_{1:T})\\ P(y_t \mid \vec{x}_{1:T}) =& \textsc{SoftMax}(\Theta \vec{z}_t).\end{aligned}$$

- **Pretraining**: the embedding function is learned from a language modeling objective on unlabeled data.
- **Direct transfer**: the embedding function is fixed, and only Θ is learned from labeled data.
- **Fine-tuning**: both the embedding function and Θ are updated during training.

Learning settings for BERT-based tagging

$$
\begin{aligned}
\vec{z}_{1:T} =& \textsc{Embed}(\vec{x}_{1:T}) \\
P(y_t \mid \vec{x}_{1:T}) =& \textsc{SoftMax}(\Theta \vec{z}_t).
\end{aligned}
$$

- **Pretraining**: the embedding function is learned from a language modeling objective on unlabeled data.
- **Direct transfer**: the embedding function is fixed, and only Θ is learned from labeled data.
- **Fine-tuning**: both the embedding function and Θ are updated during training.

Domain adaptation for BERT-based tagging

In unsupervised domain adaptation, the goal is to adapt to a new domain (such as EME), using only unlabeled data.

We propose **AdaptaBERT**:

1. Download pretrained BERT embeddings (trained on contemporary English)
2. Fine-tune EMBED using language modeling objective on unlabeled target domain text;
3. Fine-tune EMBED and Θ using tagging objective on labeled source domain text.

Summary of methods

	Source	Target
Language modeling	pretraining	
Tagging $\rightarrow \Theta$	direct transfer	
Tagging $\rightarrow$ EMBED	task tuning	

For evaluation, we map the PCHE tags to coarse-grained PTB tags (first letter only).[27]

[27] Moon and Baldridge 2007.

Summary of methods

	Source	Target
Language modeling	pretraining	AdaptaBERT
Tagging $\rightarrow \Theta$	direct transfer	
Tagging $\rightarrow$ EMBED	task tuning	

For evaluation, we map the PCHE tags to coarse-grained PTB tags (first letter only).[27]

[27] Moon and Baldridge 2007.

Summary of methods

	Source	Target
Language modeling	pretraining	AdaptaBERT
Tagging $\rightarrow \Theta$	direct transfer	supervised adaptation
Tagging $\rightarrow$ Embed	task tuning	supervised adaptation

For evaluation, we map the PCHE tags to coarse-grained PTB tags (first letter only).[27]

[27] Moon and Baldridge 2007.

Tagging results

	Early Modern English			PTB
System	Acc.	In-voc	OOV	Accuracy
Unsupervised domain adaptation				
1. Direct Transfer	77.7	83.7	61.0	91.4
2. Task tuning	85.3	90.4	71.1	98.2
3. **AdaptaBERT** (this work)	**89.8**	**90.8**	**86.8**	**98.2**
Supervised in-domain training				
4. Task tuning	98.8	99.0†	93.2†	92.4

[28] Yang and Eisenstein 2016.

Tagging results

	Early Modern English			PTB
System	Acc.	In-voc	OOV	Accuracy
Unsupervised domain adaptation				
1. Direct Transfer	77.7	83.7	61.0	91.4
2. Task tuning	85.3	90.4	71.1	98.2
3. **AdaptaBERT** (this work)	**89.8**	**90.8**	**86.8**	**98.2**
Supervised in-domain training				
4. Task tuning	98.8	99.0†	93.2†	92.4

- Task-tuned BERT outperforms the best prior work.[28]

[28] Yang and Eisenstein 2016.

Tagging results

	Early Modern English			PTB
System	Acc.	In-voc	OOV	Accuracy
Unsupervised domain adaptation				
1. Direct Transfer	77.7	83.7	61.0	91.4
2. Task tuning	85.3	90.4	71.1	98.2
3. **AdaptaBERT** (this work)	**89.8**	**90.8**	**86.8**	**98.2**
Supervised in-domain training				
4. Task tuning	98.8	99.0†	93.2†	92.4

- Task-tuned BERT outperforms the best prior work.[28]
- AdaptaBERT yields 15% improvement on OOV terms.

[28] Yang and Eisenstein 2016.

Tagging results

System	Early Modern English			PTB
	Acc.	In-voc	OOV	Accuracy
Unsupervised domain adaptation				
1. Direct Transfer	77.7	83.7	61.0	91.4
2. Task tuning	85.3	90.4	71.1	98.2
3. **AdaptaBERT** (this work)	**89.8**	**90.8**	**86.8**	**98.2**
Supervised in-domain training				
4. Task tuning	98.8	99.0†	93.2†	92.4

- Task-tuned BERT outperforms the best prior work.[28]
- AdaptaBERT yields 15% improvement on OOV terms.
- Source domain performance is not impacted.

[28] Yang and Eisenstein 2016.

Tagging results

System	Early Modern English			PTB
	Acc.	In-voc	OOV	Accuracy
Unsupervised domain adaptation				
1. Direct Transfer	77.7	83.7	61.0	91.4
2. Task tuning	85.3	90.4	71.1	98.2
3. **AdaptaBERT** (this work)	**89.8**	**90.8**	**86.8**	**98.2**
Supervised in-domain training				
4. Task tuning	98.8	99.0†	93.2†	92.4

- Task-tuned BERT outperforms the best prior work.[28]
- AdaptaBERT yields 15% improvement on OOV terms.
- Source domain performance is not impacted.
- Still no substitute for in-domain labeled data.

[28] Yang and Eisenstein 2016.

Error analysis

Most of the remaining errors are in-vocabulary, and are attributable to annotation differences on common words:

- In PTB, to gets a special tag TO, but in PCHE, the infinitival and prepositional uses are distinguished and mapped to different PTB tags (TO and IN).
- In PCHE, all is tagged as a quantifier, which is mapped to adjective; however, in PTB such usages are tagged as determiners.
- In PTB, that is tagged WDT, but in PCHE complementizers get a special tag, mapped to IN.

Such annotation differences are outside the scope of unsupervised domain adaptation.

Other work on language change

- **Constraints**: what changes are possible?[29]
- **Transition**: how does a change propagate in a community of speakers?[30]
- **Embedding**: what implications does a change have for the larger linguistic system?[31]
- **Evaluation**: what is the social meaning of a particular change?[32]
- **Actuation**: why this change, and why now?

[29] Eisenstein 2015.

[30] Eisenstein, O'Connor, et al. 2014; Goel et al. 2016.

[31] Pavalanathan and Eisenstein 2016.

[32] Pavalanathan and Eisenstein 2015.

Other work on language change

- **Constraints**: what changes are possible?[29]
- **Transition**: how does a change propagate in a community of speakers?[30]
- **Embedding**: what implications does a change have for the larger linguistic system?[31]
- **Evaluation**: what is the social meaning of a particular change?[32]
- **Actuation**: why this change, and why now?

[29] Eisenstein 2015.

[30] Eisenstein, O'Connor, et al. 2014; Goel et al. 2016.

[31] Pavalanathan and Eisenstein 2016.

[32] Pavalanathan and Eisenstein 2015.

Other work on language change

- **Constraints**: what changes are possible?[29]
- **Transition**: how does a change propagate in a community of speakers?[30]
- **Embedding**: what implications does a change have for the larger linguistic system?[31]
- **Evaluation**: what is the social meaning of a particular change?[32]
- **Actuation**: why this change, and why now?

Future work: syntactic, morphological, and phonological change; generalization beyond English; and linking language change to ongoing social changes.

---

[29] Eisenstein 2015.

[30] Eisenstein, O'Connor, et al. 2014; Goel et al. 2016.

[31] Pavalanathan and Eisenstein 2016.

[32] Pavalanathan and Eisenstein 2015.

Conclusions

- While language change poses problems for language technology, it offers new opportunities for computational social science and the study of science.
- Understanding and managing digital online discourse requires making inferences about language change.
- These research problems will require new syntheses between natural language processing, linguistics, and quantitative social science.

References I

- Altmann, Eduardo G., Janet B. Pierrehumbert, and Adilson E. Motter (2011). "Niche as a determinant of word fate in online groups". In: *PloS one* 6.5, e19009.
- Bakshy, Eytan et al. (2012). "The role of social networks in information diffusion". In: *Proceedings of the Conference on World-Wide Web (WWW)*, pp. 519–528.
- Chandrasekharan, Eshwar et al. (2018). "You Can't Stay Here: The Effectiveness of Reddit's 2015 Ban Through the Lens of Hate Speech". In: *Proceedings of Computer-Supported Cooperative Work (CSCW)*.
- Degaetano-Ortlieb, Stefania (2018). "Stylistic variation over 200 years of court proceedings according to gender and social class". In: *Proceedings of the Second Workshop on Stylistic Variation*, pp. 1–10.

References II

Devlin, Jacob et al. (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805.

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.

– (2015). "Systematic patterning in phonologically-motivated orthographic variation". In: *Journal of Sociolinguistics* 19 (2), pp. 161–188.

Eisenstein, Jacob, Amr Ahmed, and Eric P. Xing (2011). "Sparse Additive Generative Models of Text". In: *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 1041–1048.

References III

Eisenstein, Jacob, Brendan O'Connor, et al. (Nov. 2014). "Diffusion of Lexical Change in Social Media". In: *PLoS ONE* 9.

Garley, Matt and Julia Hockenmaier (2012). "Beefmoves: dissemination, diversity, and dynamics of English borrowings in a German hip hop forum". In: *Proceedings of the Association for Computational Linguistics (ACL)*, pp. 135–139.

Goel, Rahul et al. (Nov. 2016). "The Social Dynamics of Language Change in Online Networks". In: *The International Conference on Social Informatics (SocInfo)*.

References IV

Hamilton, William L, Jure Leskovec, and Dan Jurafsky (2016a). "Cultural shift or linguistic drift? comparing two computational measures of semantic change". In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*. Vol. 2016. NIH Public Access, p. 2116.

– (2016b). "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change". In: *Proceedings of the Association for Computational Linguistics (ACL)*.

Han, Xiaochuang and Jacob Eisenstein (2019). "Unsupervised Domain Adaptation of Contextualized Embeddings: A Case Study in Early Modern English". In: *arXiv preprint arXiv:1904.02817*.

References V

- Hirano, Keisuke and Guido W Imbens (2004). "The propensity score with continuous treatments". In: *Applied Bayesian modeling and causal inference from incomplete-data perspectives* 226164, pp. 73–84.
- Kroch, Anthony, Beatrice Santorini, and Ariel Diertani (2004). *Penn-Helsinki Parsed Corpus of Early Modern English*. http://www.ling.upenn.edu/hist-corpora/PPCEME-RELEASE-2/index.html.
- Kulkarni, Vivek et al. (2015). "Statistically Significant Detection of Linguistic Change". In: *Proceedings of the Conference on World-Wide Web (WWW)*, pp. 625–635.
- Mikolov, Tomas et al. (2013). "Distributed representations of words and phrases and their compositionality". In: *Advances in Neural Information Processing Systems*, pp. 3111–3119.

References VI

- Moon, Taesun and Jason Baldridge (2007). "Part-of-Speech Tagging for Middle English through Alignment and Projection of Parallel Diachronic Texts". In: *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*, pp. 390–399.
- Muralidharan, Aditi and Marti A Hearst (2013). "Supporting exploratory text analysis in literature study". In: *Literary and linguistic computing* 28.2, pp. 283–295.
- Pavalanathan, Umashanthi and Jacob Eisenstein (May 2015). "Audience-modulated variation in online social media". In: *American Speech* 90.2.
- – (Nov. 2016). "More emojis, less :) The Competition for Paralinguistic Functions in Microblog Writing". In: *First Monday* 22.11.

References VII

Peters, Matthew E et al. (2018). "Deep contextualized word representations". In: *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Pierrehumbert, Janet B. (2010). "The dynamic lexicon". In: *Handbook of Laboratory Phonology*. Ed. by A. Cohn, M. Huffman, and C. Fougeron. Oxford University Press, pp. 173–183.

Romero, Daniel M., Brendan Meeder, and Jon Kleinberg (2011). "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter". In: *Proceedings of the Conference on World-Wide Web (WWW)*, pp. 695–704.

Rosenbaum, Paul R and Donald B Rubin (1983). "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1, pp. 41–55.

References VIII

Rosenfeld, Alex and Katrin Erk (2018). "Deep Neural Models of Semantic Shift". In: *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 474–484.

Rotabi, Rahmtin, Cristian Danescu-Niculescu-Mizil, and Jon Kleinberg (2017). "Competition and selection among conventions". In: *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 1361–1370.

Soni, Sandeep, Kristina Lerman, and Jacob Eisenstein (2019). "Quantifying Semantic Progressiveness of Documents". In: *submitted to ACL*.

Stewart, Ian and Jacob Eisenstein (2018). "Making "fetch" happen: The influence of social and linguistic context on the success of lexical innovations". In: *Proceedings of Empirical Methods for Natural Language Processing (EMNLP)*.

References IX

Tan, Chenhao and Lillian Lee (2015). "All who wander: On the prevalence and characteristics of multi-community engagement". In: *Proceedings of the Conference on World-Wide Web (WWW)*, pp. 1056–1066.

Traugott, Elizabeth Closs and Richard B Dasher (2001). *Regularity in semantic change*. Cambridge University Press.

Vuillemot, Romain et al. (2009). "What's being said near "Martha"? Exploring name entities in literary text collections". In: *Symposium on Visual Analytics Science and Technology*. IEEE, pp. 107–114.

Weinreich, Uriel, William Labov, and Marvin Herzog (1968). "Empirical foundations for a theory of language change". In: *Directions for historical linguistics*. Ed. by W. P. Lehmann and Y. Malkiel. University of Texas Press, pp. 97–188.

References X

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).*

Scientific abstracts

Predictors	M1	M2	M3	M4
Intercept	1.7929 (0.0025)	1.7964 (0.0026)	1.6389 (0.0027)	1.4181 (0.0031)
Out degree	0.0166 (0.0000)	0.0166 (0.0000)	0.0165 (0.0000)	0.0162 (0.0000)
Age	0.0863 (0.0001)	0.0863 (0.0001)	0.0933 (0.0001)	0.0973 (0.0001)
Length	0.0047 (0.0000)	0.0047 (0.0000)	0.0045 (0.0000)	0.0047 (0.0000)
No. of Authors	0.0406 (0.0002)	0.0406 (0.0002)	0.0418 (0.0002)	0.0421 (0.0002)
BoWs		0.0000 (0.0000)	0.0000 (0.0000)	0.0000 (0.0000)
Progressiveness			0.0138 (0.0001)	
Prog. Q2				0.1876 (0.0021)
Prog. Q3				0.4200 (0.0023)
Prog. Q4				0.5862 (0.0023)
Log Likelihood	-3085945	-3085891	-3057184	-3050474

Legal opinions

Predictors	M1	M2	M3	M4
Intercept	1.8171 (0.0051)	1.9246 (0.0053)	1.9210 (0.0055)	1.6911 (0.0081)
Out degree	0.0150 (0.0001)	0.0089 (0.0002)	0.0088 (0.0002)	0.0086 (0.0002)
Age	0.0155 (0.0001)	0.0140 (0.0001)	0.0141 (0.0001)	0.0156 (0.0001)
Length	0.0003 (0.0000)	0.0004 (0.0000)	0.0004 (0.0000)	0.0004 (0.0000)
BoWs		0.0002 (0.0000)	0.0002 (0.0000)	0.0002 (0.0000)
Progressiveness			0.0002 (0.0001)	
Prog. Q2				0.2007 (0.0079)
Prog. Q3				0.2566 (0.0082)
Prog. Q4				0.3336 (0.0082)
Log Likelihood	-231778	-228538	-228535	-227663