Human-Centered Natural Language Processing

CSE 354
The “Task” of human-centered NLP

Most NLP Tasks. E.g.
- POS Tagging
- Document Classification
- Sentiment Analysis
- Stance Detection
- Mental Health Risk Assessment
- …
  (language modeling, QA, …)
The “Task” of human-centered NLP

Most NLP Tasks. E.g.

- POS Tagging
- Document Classification
- Sentiment Analysis
- Stance Detection
- Mental Health Risk Assessment
- …
- (language modeling, QA, …)
The “Task” of human-centered NLP

Most NLP Tasks. E.g.
- POS Tagging
- Document Classification
- Sentiment Analysis
- Stance Detection
- Mental Health Risk Assessment
- …
  (language modeling, QA, …)

How to include extra-linguistics?
- Additive Inclusion
- Adaptive Extralinguistics
  - Adapting Embeddings
  - Adapting Models
- Correcting for bias

age
gender
personality
expertise
beliefs
…
Natural Language Processing

Psychological & Health Sciences
Problem

Natural language is written by
Problem

Natural language is written by people.
Problem

Natural language is written by people.

That’s sick
Problem

Natural language is written by people.

That's sick
Natural language is generated by people.

People have different beliefs, backgrounds, styles, vocabularies, preferences, knowledge, personalities, …
Natural language is generated by people.

People have different beliefs, backgrounds, styles, vocabularies, preferences, knowledge, personalities, …, and our language reflects these differences.
Natural language is generated by people.

People have different beliefs, backgrounds, styles, vocabularies, preferences, knowledge, personalities, …,

and our language reflects these differences.
Natural language is generated by people.

- Shannon, 1948
- Mosteller & Wallace, 1963
- Clark & Schober, 1992
- Mairesse, Walker, et al., 2007
- Hovy & Soegaard, 2015
Natural language is generated by people.

“The common misconception is that language has got to do with words and what they mean. It does not. It has to do with people and what they mean.”

- Shannon, 1948
- Mosteller & Wallace, 1963
- Clark & Schober, 1992
- Mairesse, Walker, et al., 2007
- Hovy & Soegaard, 2015
Natural language is generated by people.

(MOST LINGUISTS)

Shannon, 1948
Mosteller & Wallace, 1963
Clark & Schober, 1992
Mairesse, Walker, et al., 2007
Hovy & Soogaard, 2015
Natural language is generated by people.

Yet, our models:
Natural language is generated by people.

Yet, our models:

e.g. Document Classification: Stance

Natural language is generated by people.

Yet, our models:

- e.g. Document Classification: Stance
  - e.g., pro gun control? yes, no

Natural language is generated by people.

- Document Classification: Stance
  - e.g., pro gun control?
  - yes, no

Natural language is generated by people.

E.g., Document Classification: Stance

E.g., pro gun control? yes, no

Natural language is generated by people.

- personality
- demographics
- emotional states
- political ideology
- linguistic style (Pennebaker, 2007)
- latent user traits (Kulkarni et al., 2018)

E.g. Document Classification: Stance

E.g. pro gun control? yes, no

Natural language is generated by people.

What this means for NLP:

1. Our data are inherently multi-level.
2. Often, there are "already-available" human attributes.
3. Our data and models are (human) biased.
Natural language is generated by people.

What this means for NLP:

1. Our data are inherently multi-level.
2. Often, there are "already-available" human attributes.
3. Our data and models are (human) biased.
Natural language is generated by people.

What this means for NLP:

1. Our data are inherently multi-level.
2. Often, there are "already-available" human attributes.
3. Our data and models are (human) biased.
Approaches to Human Factor Inclusion

1. Adaptive: Allow meaning if language to change depending on human context. (also called “compositional”)
   (e.g. “sick” said from a young individual versus old individual)
Approaches to Human Factor Inclusion

1. Adaptive: Allow meaning of language to change depending on human context. (also called “compositional”)
   (e.g. “sick” said from a young individual versus old individual)

   (e.g. age and distinguishing PTSD from Depression)
Approaches to Human Factor Inclusion

1. **Adaptive**: Allow meaning if language to change depending on human context. (also called “compositional”)  
   (e.g. “sick” said from a young individual versus old individual)

2. **Additive**: Include direct effect of human factor on outcome.  
   (e.g. age and distinguishing PTSD from Depression)

3. **Bias Correction**: Optimize so as not to pick up on unwanted relationships.  
   (e.g. image captioner label pictures of men in kitchen as women)
Approaches to Human Factor Inclusion

1. Adaptive: Allow meaning of language to change depending if language to change depending on human context. (also called “compositional”)
   (e.g. “sick” said from a young individual versus old individual)

   (e.g. age and distinguishing PTSD from Depression)

3. Bias Correction: Optimize so as not to pick up on unwanted relationships.
   (e.g. image captioner label pictures of men in kitchen as women)
Adaptation Approach: Domain Adaptation

Features for: source \[ \Phi^s(x) = \langle x, x, 0 \rangle \] \quad \text{target} \quad \Phi^t(x) = \langle x, 0, x \rangle

Frustratingly Easy Domain Adaptation

Hal Daumé III
School of Computing
University of Utah
Salt Lake City, Utah 84112
me@hal3.name

Abstract

We describe an approach to domain adaptation that is appropriate exactly in the unsupervised case. The fully supervised case models the following scenario. We have access to a large, annotated corpus of data for...
Adaptation Approach: Domain Adaptation

Features for: source \( \Phi^s(\mathbf{x}) = \langle \mathbf{x}, \mathbf{x}, 0 \rangle \), \( \Phi^s(\mathbf{x}) = \langle \mathbf{x}, 0, \mathbf{x} \rangle \)

target

newX = []
for all \( x \) in source\_x:
    newX.append(\( x + x + [0]^{\text{len}(x)} \))
for all \( x \) in target\_x:
    newX.append(\( x + [0]^{\text{len}(x)}, x \))

Frustratingly Easy Domain Adaptation

Hal Daumé III
School of Computing
University of Utah
Salt Lake City, Utah 84112
me@halj3.name

Abstract
We describe an approach to domain adaptation that is appropriate exactly in the setting of our thesis. The fully supervised case models the following scenario. We have access to a large, annotated corpus of data for the source domain. Supervised case. The fully supervised case models the following scenario. We have access to a large, annotated corpus of data for the source domain.
Adaptation Approach: Domain Adaptation

Features for: source | target
---|---
\( \Phi^s(x) = \langle x, x, 0 \rangle \), \( \Phi^t(x) = \langle x, 0, x \rangle \)

\[
\text{newX} = []
\]

for all \( x \) in source_x:
newX.append(\( x + x + [0] \times \text{len}(x) \))

for all \( x \) in target_x:
newX.append(\( x + [0] \times \text{len}(x), x \))

\[
\text{newY} = \text{source}_y + \text{target}_y
\]

\[
\text{model} = \text{model}.\text{train} (\text{newX}, \text{newY})
\]

Frustratingly Easy Domain Adaptation

Hal Daumé III
School of Computing
University of Utah
Salt Lake City, Utah 84112
me@halj3.name

Abstract

We describe an approach to domain adaptation that is appropriate exactly in the cases of interest. The fully supervised case models the following scenario. We have access to a large, annotated corpus of data from the source domain. However, we do not have access to labeled examples from the target domain. Instead, we have access to an unannotated corpus from the target domain. Our approach is to train a model on the source data and then fine-tune it on the target data. This allows us to leverage the large amount of labeled data from the source domain to improve performance on the target domain.
Human Factors

--- Any attribute, represented as a continuous or discrete variable, of the humans generating the natural language.

E.g.

- Gender
- Age
- Personality
- Ethnicity
- Socio-economic status
Adaptation Approach: Factor Adaptation

Human Centered NLP with User-Factor Adaptation

Vera L. Lyan, Youngsoo Son, Vivek Kulkarni, Niranjan Balasubramanian, and H. Andrew Schwartz
Stony Brook University
Stony Brook, NY
{velyan, yson, vvkulkarni, nirajan, has}@cs.stonybrook.edu

Residualized Factor Adaptation
for Community Social Media Prediction Tasks

Mohammadzaman Zaman, H. Andrew Schwartz, Veronica E. Lyan,
Salvatore Giorgi, and Niranjan Balasubramanian

Abstract

We pose the general task of user-factor adaptation — adapting supervised learning models to real-valued user factors that are either background or domain-specific — as a factor analysis problem. This approach is motivated by previous work on NLP and related fields. By modeling the interaction between user factors and model parameters, we are able to improve the performance of the model on real-world datasets.

Residualized Factor Adaptation
for Community Social Media Prediction Tasks

Mohammadzaman Zaman, H. Andrew Schwartz, Veronica E. Lyan,
Salvatore Giorgi, and Niranjan Balasubramanian

Abstract

Predictive models for social media language processing in capturing community linked to socio-demographic factors (age, gender, race, education, income levels) with many social scientific studies supporting their predictive power. This work aims to enhance the performance of these models by applying factor adaptation techniques.
Adaptation

typically requires putting people into discrete bins
“most latent variables of interest to psychiatrists and personality and clinical psychologists are dimensional [continuous]”
(Haslam et al., 2012)
“most latent variables of interest to psychiatrists and personality and clinical psychologists are dimensional [continuous]”

(Haslam et al., 2012)
“most latent variables of interest to psychiatrists and personality and clinical psychologists are dimensional [continuous]”
(Haslam et al., 2012)
Our Method: Continuous Adaptation

(User Factors | Train Instances | Labels)
-0.2  
0.6  
0.3  
-0.4  

(Continuous Adaptation)  

(Transformed Instances | Labels)

(Lynn et al., 2017)
Our Method: Continuous Adaptation

- User Factors
  - -.2
  - .6
  - .3
  - -.4

- Train Instances

- Labels

- Continuous Adaptation

- Transformed Instances

- Labels

Gender Score
-.2

Features
X

Original
X

(Lynn et al., 2017)
Our Method: Continuous Adaptation

User Factors | Train Instances | Labels
---|---|---
-.2 | | |
.6 | | |
.3 | | |
-.4 | | |

Continuous Adaptation

Transformed Instances | Labels
---|---
| |
| |
| |
| |

Gender Score
-.2

Features
X

Original
X

Gender Copy
compose(-.2, X)

(Lynn et al., 2017)
User Factor Adaptation: Handling multiple factors

Replicate features for each factor:

A compositional function $c$ combines $d$ user factor scores $f_{u,d}$ with original feature values $x$:

$$\Phi(x, u) = \langle x, c(f_{u,1}, x), c(f_{u,2}, x), \cdots, c(f_{u,d}, x) \rangle$$

(Lynn et al., 2017)
User Factor Adaptation: Handling multiple factors

Replicate features for each factor:

A compositional function $c$ combines $d$ user factor scores $f_{ud}$ with original feature values $x$:

$$\Phi(x, u) = \langle x, c(f_{u,1}, x), c(f_{u,2}, x), \ldots, c(f_{u,d}, x) \rangle$$

<table>
<thead>
<tr>
<th>User</th>
<th>Factor Classes</th>
<th>Augmented Instance $\Phi(x, u)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>$F_1$</td>
<td>$\langle x, x, 0, 0, \ldots, 0 \rangle$</td>
</tr>
<tr>
<td>User 2</td>
<td>$F_2$</td>
<td>$\langle x, 0, x, 0, \ldots, 0 \rangle$</td>
</tr>
<tr>
<td>User 3</td>
<td>$F_1, F_3$</td>
<td>$\langle x, x, 0, x, \ldots, 0 \rangle$</td>
</tr>
<tr>
<td>User 4</td>
<td>$F_k$</td>
<td>$\langle x, 0, 0, \ldots, 0, x \rangle$</td>
</tr>
</tbody>
</table>

Table 1: Discrete Factor Adaptation: Augmentations of an original instance vector $x$ under different factor class mappings. With $k$ domains the augmented feature vector is of length $n(k + 1)$. (Lynn et al., 2017)
User Factor Adaptation: Handling multiple factors

Replicate features for each factor:

A compositional function $c$ combines $d$ user factor scores $f_{u,d}$ with original feature values $x$:

$$\Phi(x, u) = \langle x, c(f_{u,1}, x), c(f_{u,2}, x), \ldots, c(f_{u,d}, x) \rangle$$

<table>
<thead>
<tr>
<th>User</th>
<th>Factor Classes</th>
<th>Augmented Instance $\Phi(x, u)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>User 1</td>
<td>$F_1$</td>
<td>$\langle x, x, 0, 0, \ldots, 0 \rangle$</td>
</tr>
<tr>
<td>User 2</td>
<td>$F_2$</td>
<td>$\langle x, 0, x, 0, \ldots, 0 \rangle$</td>
</tr>
<tr>
<td>User 3</td>
<td>$F_1, F_3$</td>
<td>$\langle x, x, 0, x, \ldots, 0 \rangle$</td>
</tr>
<tr>
<td>User 4</td>
<td>$F_k$</td>
<td>$\langle x, 0, 0, \ldots, 0, x \rangle$</td>
</tr>
</tbody>
</table>

Table 1: Discrete Factor Adaptation: Augmentations of an original instance vector $x$ under different factor class mappings. With $k$ domains the augmented feature vector is of length $n(k + 1)$. (Lynn et al., 2017)
Main Results

Adaptation improves over unadapted baselines (Lynn et al., 2017)

<table>
<thead>
<tr>
<th>Task</th>
<th>Metric</th>
<th>No Adaptation</th>
<th>Gender</th>
<th>Personality</th>
<th>Latent (User Embed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stance</td>
<td>F1</td>
<td>64.9</td>
<td>65.1 (+0.2)</td>
<td>66.3 (+1.4)</td>
<td>67.9 (+3.0)</td>
</tr>
<tr>
<td>Sarcasm</td>
<td>F1</td>
<td>73.9</td>
<td>75.1 (+1.2)</td>
<td>75.6 (+1.7)</td>
<td>77.3 (+3.4)</td>
</tr>
<tr>
<td>Sentiment</td>
<td>Acc.</td>
<td>60.6</td>
<td>61.0 (+0.4)</td>
<td>61.2 (+0.6)</td>
<td>60.7 (+0.1)</td>
</tr>
<tr>
<td>PP-Attach</td>
<td>Acc.</td>
<td>71.0</td>
<td>70.7 (-0.3)</td>
<td>70.2 (-0.8)</td>
<td>70.8 (-0.2)</td>
</tr>
<tr>
<td>POS</td>
<td>Acc.</td>
<td>91.7</td>
<td>91.9 (+0.2)</td>
<td>91.2 (-0.5)</td>
<td>90.9 (-0.8)</td>
</tr>
</tbody>
</table>
Example: How Adaptation Helps

Women
more adjectives→sarcasm

Men
more adjectives→no sarcasm
Problem

User factors are not always available.
Solution: User Factor Inference

past tweets

Niranjan @b_niranjan · Sep 2
There must be a word for trending hashtags that you know you will regret if you click. Is there?

Niranjan @b_niranjan · Aug 31
Passwords spiral: Forget password for the acct you use twice a year. Ask for reset. Can't use previous. Create a new one to forget later.

Niranjan @b_niranjan · Jul 31
Thrilled to hear @acl2017’s diversity efforts as the first thing in the conference.

inferred factors

Known
Age (Sap et al. 2014)
Gender (Sap et al. 2014)
Personality (Park et al. 2015)

Latent
User Embeddings (Kulkarni et al. 2017)
Word2Vec
TF-IDF
Using more background tweets to infer factors produces larger gains

**Background Size**

- **personality (cont)**
- **user embed (cont)**

- **gains over baseline (f1)**
- **background size**
**Full User Factors Adaptation Pipeline: with latent factors from training**

<table>
<thead>
<tr>
<th>doc-id</th>
<th>user-id</th>
<th>document</th>
<th>embdngs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>42</td>
<td>text...</td>
<td>embdngs</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>text...</td>
<td>embdngs</td>
</tr>
<tr>
<td>3</td>
<td>16</td>
<td>text...</td>
<td>embdngs</td>
</tr>
<tr>
<td>4</td>
<td>42</td>
<td>text...</td>
<td>embdngs</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>text...</td>
<td>embdngs</td>
</tr>
<tr>
<td>6</td>
<td>16</td>
<td>text...</td>
<td>embdngs</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Total documents**: d = 128
Full User Factors Adaptation Pipeline: with latent factors from training

Step 1: Create User Factors
**Full User Factors Adaptation Pipeline:** with latent factors from training

**Step 2: Create User-adapted Features**

- **Document Text**: text...
- **User-adapted Embeddings**: embdngs
- **User x Factors**: Averages of embdngs
- **PCA**: Reduces dimensionality to d = 128
- **User x avg_embeddings**
  
  - **Users x avg_embeddings**
  - **PCA**
  - **d = 128**

- **User-adapted Embeddings**: emb x f1; emb x f2; emb x f3
- **User x Factors**: 42, 16, 12

- **Total Documents**: doc-id, user-id

- **User-adapted Embeddings**: 1, 2, 3, 4, ...

- **Dimensions**: d = 128 and lower dimensions (d = 3)
Full User Factors Adaptation Pipeline: with latent factors from training

Step 3: Train Model

Then feed these as features into your document level classifier or regressor.
Full User Factors Adaptation Pipeline: with latent factors from training

This was training data; now assume test

What about when predicting on new documents?

What about when predicting on new documents?
What about when predicting on new documents? (easy as A, B, C)

A. Save the transformation (V) from PCA during training

**Full User Factors Adaptation Pipeline: with latent factors from training**

- **User-Adapted Embeddings**
  - User-id
  - Text...
  - Embdings
  - Avg

- **Document**
  - Doc-id
  - User-id
  - Text...
  - Embdings
  - Avg

- **Total Documents**
  - Doc-id
  - User-id
  - Text...
  - Embdings
  - Avg

- **User x Factors**
  - User-id
  - Factors
  - N users

- **PCA**
  - Transformation Matrix (V)
  - D = 128

- **Users x Avg_embeddings**
  - N users
  - Avg

- **What about when predicting on new documents?**

**Note:** This was training data; now assume test
Full User Factors Adaptation Pipeline: with latent factors from training

What about when predicting on new documents?
(easy as A, B, C)

A. Save the transformation (V) from PCA during training
B. Apply V to user x avg_embeddings matrix during test/trial.

...
Full User Factors Adaptation Pipeline: with latent factors from training

What about when predicting on new documents? (easy as A, B, C)

A. Save the transformation (V) from PCA during training
B. Apply V to user x avg_embeddings matrix during test/trial.
C. Adapt document features by user factors just like in training.
Approaches to Human Factor Inclusion

1. **Adaptive**: Allow meaning if language to change depending on human context. (also called “compositional”)
   (e.g. “sick” said from a young individual versus old individual)

2. **Additive**: Include direct effect of human factor on outcome.
   (e.g. age and distinguishing PTSD from Depression)

3. **Bias Correction**: Optimize so as not to pick up on unwanted relationships.
   (e.g. image captioner label pictures of men in kitchen as women)
Approaches to Human Factor Inclusion

1. Adaptive: Allow meaning if language to change depending on human context. (also called “compositional”)  
   (e.g. “sick” said from a young individual versus old individual)

   (e.g. age and distinguishing PTSD from Depression)

3. Bias Correction: Optimize so as not to pick up on unwanted relationships.  
   (e.g. image captioner label pictures of men in kitchen as women)
Ethics in NLP: Human Factor Inclusion

1. Adaptive: Allow meaning if language to change depending on human context. (also called “compositional”) (e.g. “sick” said from a young individual versus old individual)

2. Additive: Include direct effect of human factor on outcome. (e.g. age and distinguishing PTSD from Depression)

3. Bias Correction: Optimize so as not to pick up on unwanted relationships. (e.g. image captioner label pictures of men in kitchen as women)
Ethics in NLP

Bias

Privacy

Ethical Research
Ethics in NLP

Types of bias in NLP tasks:

- **Outcome Disparity**: Predicted distribution given A, are dissimilar from ideal distribution given A
  - Selection bias
  - Label bias
  - Over-amplification

- **Error Disparity**: Predicts less accurate for authors of given demographics.

- **Semantic Bias**: Representations of meaning store demographic associations.

Two Examples

The WSJ Effect

Jørgensen/Hovy/Søgaard, 2015
Hovy & Søgaard, 2015

model accuracy

distance from “standard” WSJ author demographics
Two Examples

The WSJ Effect

Jørgensen/Hovy/Søgaard, 2015
Hovy & Søgaard, 2015

model accuracy

distance from “standard” WSJ author demographics
Two Examples

Two Examples

Our data and models are (human) biased.

“Outcome Disparity”

“Error Disparity”

Person-level
- attribute = 1
- attribute = 2
Our data and models are (human) biased.

**Outcome Disparity**

**Error Disparity**
Our data and models are (human) biased.

“Outcome Disparity”

“Error Disparity”
outcome disparity
The distribution of outcomes, given attribute $A$, is dissimilar than the ideal distribution:
$$Q(\hat{Y}_t | A) \neq P(Y_t | A)$$

error disparity
The distribution of error ($\epsilon$) over at least two different values of an attribute ($A$) are unequal:
$$Q(\epsilon_i | A_i) \neq Q(\epsilon_i | A_j)$$

Target Population
(features $X_{target}$) predict biased outcomes $\hat{Y}_{target}$)

The distribution of outcomes, given attribute $A$, is dissimilar than the ideal distribution:

$$Q(\hat{Y}_t | A) \neq P(Y_t | A)$$

The distribution of error ($\epsilon$) over at least two different values of an attribute ($A$) are unequal:

$$Q(\epsilon_t | A_i) \neq Q(\epsilon_t | A_j)$$
predictive biases in natural language processing models: a conceptual framework and overview.

source population
(features x_{source} \rightarrow \textit{fit} \rightarrow \textit{outcomes} y_{source})

target population
(features x_{target} \rightarrow \textit{predict} \rightarrow \textit{biased outcomes} \hat{y}_{target})

outcome disparity
The distribution of outcomes, given attribute A, is dissimilar than the ideal distribution:
Q(\hat{Y}_t|A) \neq P(Y_t|A)

error disparity
The distribution of error (\epsilon) over at least two different values of an attribute (A) are unequal:
Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j)

label bias
Biased annotations, interaction, or latent bias from past classifications.

Biased annotations, interaction, or latent bias from past classifications.

The distribution of outcomes, given attribute $A$, is dissimilar than the ideal distribution:

$$Q(\hat{Y}_t|A) \neq P(Y_t|A)$$

The distribution of error ($\epsilon$) over at least two different values of an attribute ($A$) are unequal:

$$Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j)$$

The sample of observations themselves are not representative of the application population.

$$\theta_{embedding}$$

$X_{source}$ (Model Side) $\rightarrow$ $Y_{source}$

$X_{target}$ (Application Side) $\rightarrow$ $\hat{Y}_{target}$

Biased annotations, interaction, or latent bias from past classifications.

The model discriminates on a given human attribute beyond its source base-rate.

The distribution of outcomes, given attribute $A$, is dissimilar than the ideal distribution:

\[ Q(\hat{Y}_t|A) \neq P(Y_t|A) \]

The distribution of error ($\epsilon$) over at least two different values of an attribute ($A$) are unequal:

\[ Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j) \]

The sample of observations themselves are not representative of the application population.
features
source
X
features
target
label bias
label bias
Biased annotations, interaction, or latent bias from past classifications.
outcome disparity
The distribution of outcomes, given attribute $A$, is dissimilar than the ideal distribution:
$Q(\hat{Y}_t|A) \neq P(Y_t|A)$

selection bias
The sample of observations themselves are not representative of the application population.
error disparity
The distribution of error ($\epsilon$) over at least two different values of an attribute ($A$) are unequal:
$Q(\epsilon_t|A_i) \neq Q(\epsilon_t|A_j)$

Semantics bias
Non-ideal associations between attributed lexeme (e.g. gendered pronouns) and non-attributed lexeme (e.g. occupation).

potential origin
over-amplification
The model discriminates on a given human attribute beyond its source base-rate.

Embedding Corpus
(features $\theta_{embedding}$)
(Source Population
(features $X_{source}$)
(Source Population
(outcomes $Y_{source}$)
(Target Population
(features $X_{target}$)
(Target Population
(biased outcomes $\hat{Y}_{target}$)

E.g. Coreference resolution: connecting entities to references (i.e. pronouns).

“The doctor told Mary that she had run some blood tests.”

**semantic bias**
Non-ideal associations between attributed lexeme (e.g. gendered pronouns) and non-attributed lexeme (e.g. occupation).

**selection bias**
The sample of observations themselves are not representative of the application population.

**error disparity**
The distribution of error ($\epsilon$) over at least two different values of an attribute ($A$) are unequal:

$$Q(\epsilon_i | A_i) \neq Q(\epsilon_i | A_j)$$

Ethics in NLP

Types of bias in NLP tasks:

● **Outcome Disparity:** Predicted distribution given A, are dissimilar from ideal distribution given A
  ○ Selection bias
  ○ Label bias
  ○ Over-amplification

● **Error Disparity:** Predicts less accurate for authors of given demographics.

● **Semantic Bias:** Representations of meaning store demographic associations.

Ethics in NLP

Types of bias in NLP tasks:

- **Predictive Bias**: Predicted distribution given A, are dissimilar from ideal distribution given A
  - Selection bias
  - Label bias
  - Over-amplification

- **Bias in Error**: Predicts less accurate for authors of given demographics.

- **Semantic Bias**: Representations of meaning store demographic associations.

E.g. Coreference resolution: connecting entities to references (i.e. pronouns).

“The doctor told Mary that she had run some blood tests.”

- **Semantic Bias**: Representations of meaning store demographic associations.

Work in progress; Hovy et al., 2019
Ethics in NLP

Privacy

- Risk Categories:
  - Revealing unintended private information
  - Targeted persuasion
Ethics in NLP

Privacy

- Risk Categories:
  - Revealing unintended private information
  - Targeted persuasion
- Mitigation strategies:
  - Informed consent -- let participants know
  - Do not share / secure storage
  - Federated learning -- separate and obfuscate to the point of preserving privacy
  - Transparency in information targeting
    “You are being shown this ad because ...”
Ethics in NLP

Human Subjects Research

Observational versus Interventional
Ethics in NLP

Human Subjects Research

Observational versus Interventional

(The Belmount Report, 1979)

(i) Distinction of research from practice.
(ii) Risk-Benefit criteria
(iii) Appropriate selection of human subjects for participation in research
(iv) Informed consent in various research settings.
Natural language is generated by people.

What this means for NLP:

1. Our data are inherently multi-level.
2. Often, there are "already-available" human attributes.
3. Our data and models are (human) biased.
Approaches to Human Factor Inclusion

1. Adaptive: Allow meaning if language to change depending on human context. (also called “compositional”) (e.g. “sick” said from a young individual versus old individual)

2. Additive: Include direct effect of human factor on outcome. (e.g. age and distinguishing PTSD from Depression)

3. Bias Correction: Optimize so as not to pick up on unwanted relationships. (e.g. image captioner label pictures of men in kitchen as women)
Example 1: Individual Heart Disease
Example 2: Twitter Language + Socioeconomics
Additive (Residualized Control)
Additive (Residualized Control)

Challenges:

- High-dimensional, sparse, and noisy.
- few and well estimated

language

controls
Additive (Residualized Control)

Effectively use both low dimensional control features and high-dimensional, noisy language features:

1. **Train a control model** using the control values
2. **Calculate the residual** error and consider it as the new label
3. **Train a language model** over the new labels
Additive (Residualized Control)

Residualize control (additive model):

- Extra-linguistics
- Language

1. Predictive Analytics
2. Error

- Prediction

Adaptive model:

- Language [L]
- Extra-linguistics [e1, e2, ..., em]

1. X
2. [e1.L, e2.L, ..., em.L]

1. Predictive Model
2. Prediction

(Zamani et al., EACL 2017)
Additive (Residualized Control)

Effectively use both low dimensional control features and high-dimensional, noisy language features:

1. **Train a control model** using the control values
2. **Calculate the residual** error and consider it as the new label
3. **Train a language model** over the new labels
Model:

\[ Y = \alpha x_1 + \beta x_2 + \gamma \]

Both learn the same linear model above, but

- Different learning algorithms per variable type.
- Different penalization methods
Residualized Control Model

Combining Adaptive and Additive

Two Goals:

1. **Adaptive**: adapt to given human attributes
   (*user factor adaptation*; Lynn, Balasubramanian, Son, Kulkarni & Schwartz, *EMNLP* 2017)

2. **Additive**: predict beyond given attributes
   (*residualized control*; Zamani & Schwartz, *EACL* 2017)
Solution: Residualized Factor Adaptation

1) Factor Selection
- Factors: $f_1, f_2, \ldots, f_i$

2) Factor Adaptation
- Language Data: $L_r, L_n$
- Adapted Language Data: $L_r, A_r, L_n, A_n$

3) Control Model
- Residual Error
- Label

4) Adapted Language Model
- Prediction
## Results: County Health Predictions

<table>
<thead>
<tr>
<th>Lang.</th>
<th>Controls Only</th>
<th>Added-Controls</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Dis</td>
<td>0.585</td>
<td>0.608</td>
</tr>
<tr>
<td>Suicide</td>
<td>0.414</td>
<td>0.431</td>
</tr>
<tr>
<td>Poor Health</td>
<td>0.602</td>
<td>0.641</td>
</tr>
<tr>
<td>Life Satis.</td>
<td>0.209</td>
<td>0.335</td>
</tr>
</tbody>
</table>

Avg. 0.453 0.440 0.503
Results: County Health Predictions

<table>
<thead>
<tr>
<th>Lang.</th>
<th>Controls Only</th>
<th>Added-Controls</th>
<th>Res-Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Dis</td>
<td>0.585</td>
<td>0.514</td>
<td>0.608</td>
</tr>
<tr>
<td>Suicide</td>
<td>0.414</td>
<td>0.307</td>
<td>0.431</td>
</tr>
<tr>
<td>Poor Health</td>
<td>0.602</td>
<td>0.609</td>
<td>0.641</td>
</tr>
<tr>
<td>Life Satis.</td>
<td>0.209</td>
<td>0.329</td>
<td>0.335</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.453</td>
<td>0.440</td>
<td>0.503</td>
</tr>
</tbody>
</table>
## Results: County Health Predictions

<table>
<thead>
<tr>
<th>Lang.</th>
<th>Controls Only</th>
<th>Added-Controls</th>
<th>Res-Control</th>
<th>FA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Dis</td>
<td>0.585</td>
<td>0.514</td>
<td>0.608</td>
<td>0.628</td>
</tr>
<tr>
<td>Suicide</td>
<td>0.414</td>
<td>0.307</td>
<td>0.431</td>
<td>0.460</td>
</tr>
<tr>
<td>Poor Health</td>
<td>0.602</td>
<td>0.609</td>
<td>0.641</td>
<td>0.661</td>
</tr>
<tr>
<td>Life Satis.</td>
<td>0.209</td>
<td>0.329</td>
<td>0.335</td>
<td>0.372</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.453</td>
<td>0.440</td>
<td>0.503</td>
<td>0.530</td>
</tr>
</tbody>
</table>
## Results: County Health Predictions

<table>
<thead>
<tr>
<th>Lang.</th>
<th>Controls Only</th>
<th>Added-Controls</th>
<th>Res-Control</th>
<th>FA</th>
<th>RFA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Dis</td>
<td>0.585</td>
<td>0.514</td>
<td>0.608</td>
<td>0.628</td>
<td>0.635</td>
</tr>
<tr>
<td>Suicide</td>
<td>0.414</td>
<td>0.307</td>
<td>0.431</td>
<td>0.460</td>
<td>0.494</td>
</tr>
<tr>
<td>Poor Health</td>
<td>0.602</td>
<td>0.609</td>
<td>0.641</td>
<td>0.661</td>
<td>0.674</td>
</tr>
<tr>
<td>Life Satis.</td>
<td>0.209</td>
<td>0.329</td>
<td>0.335</td>
<td>0.372</td>
<td>0.352</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.453</td>
<td>0.440</td>
<td>0.503</td>
<td>0.530</td>
<td>0.539</td>
</tr>
</tbody>
</table>
### Results: County Health Predictions

<table>
<thead>
<tr>
<th>Language</th>
<th>Heart Dis</th>
<th>Suicide</th>
<th>Poor Health</th>
<th>Life Satis.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Controls Only</td>
<td>0.585</td>
<td>0.414</td>
<td>0.602</td>
<td>0.209</td>
</tr>
<tr>
<td>Added Controls</td>
<td>0.514</td>
<td>0.307</td>
<td>0.609</td>
<td>0.329</td>
</tr>
<tr>
<td>Res-Control</td>
<td>0.608</td>
<td>0.431</td>
<td>0.641</td>
<td>0.335</td>
</tr>
<tr>
<td>All Factors</td>
<td>0.628</td>
<td>0.460</td>
<td>0.661</td>
<td>0.372</td>
</tr>
<tr>
<td>FA</td>
<td>0.635</td>
<td>0.494</td>
<td>0.674</td>
<td>0.352</td>
</tr>
<tr>
<td>RFA</td>
<td>0.655</td>
<td>0.510</td>
<td>0.682</td>
<td>0.396</td>
</tr>
</tbody>
</table>

**Avg.**

|            | 0.453 | 0.440 | 0.503 | 0.530 | 0.539 | 0.560 |

**Variance explained ($R^2$)**
Natural language is generated by people.

What this means for NLP:

1. Our data are inherently multi-level.
2. Often, there are "already-available" human attributes.
3. Our data and models are (human) biased.
Natural language is generated by people.

What this means for NLP:

1. Our data are inherently multi-level.
2. Often, there are "already-available" human attributes.
3. Our data and models are (human) biased.
Data are inherently multi-level.
1,639,750 tweets from 5,226 users in 420 counties
Data are inherently multi-level.

<table>
<thead>
<tr>
<th>Distribution</th>
<th>Message</th>
<th>User</th>
<th>County</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-gram</td>
<td>topic</td>
<td>Lex.</td>
</tr>
<tr>
<td>Power Law</td>
<td>.71</td>
<td>.10</td>
<td>.00</td>
</tr>
<tr>
<td>Log-Normal</td>
<td>.25</td>
<td>.89</td>
<td>1.00</td>
</tr>
<tr>
<td>Normal</td>
<td>.04</td>
<td>.01</td>
<td>.00</td>
</tr>
</tbody>
</table>

Proportion best fit by the given distribution.

Data are inherently multi-level.

Multi-level Attention and Sequence Model

Data are inherently multi-level.

<table>
<thead>
<tr>
<th>Method</th>
<th>$d$</th>
<th>O</th>
<th>C</th>
<th>E</th>
<th>A</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT + DAN (Park et al., 2015)</td>
<td>768</td>
<td>.60</td>
<td>.51</td>
<td>.54</td>
<td>.51</td>
<td>.52</td>
</tr>
<tr>
<td>Multi-level Attention</td>
<td>5106</td>
<td>.63</td>
<td>.52</td>
<td>.56</td>
<td>.54</td>
<td>.53</td>
</tr>
<tr>
<td>Multi-level Attention + Ridge</td>
<td>200</td>
<td>.63</td>
<td>.52</td>
<td>.55</td>
<td>.51</td>
<td>.54</td>
</tr>
<tr>
<td></td>
<td>5306</td>
<td>.66</td>
<td>.54</td>
<td>.58</td>
<td>.56</td>
<td>.56</td>
</tr>
</tbody>
</table>

*bold: $p < .01$*

Data are inherently multi-level.

Differential Language Analysis

Input:

- Linguistic features
- Human or community attribute

Output:

- Features distinguishing attribute

Goal: Data-driven insights about an attribute
E.g. Words distinguishing communities with increases in real estate prices.
Differential Language Analysis

Input:

Linguistic features

Human or community attribute

Output:

Features distinguishing attribute

Goal: Data-driven insights about an attribute
Differential Language Analysis

Volunteer Data

1) Linguistic feature extraction
   a) words and phrases
   b) topics

2) Correlation analysis

3) Visualization

social media messages

gender
location
age
health
...

amazing
love
everyone
excited
day
for

thanksgiving
cdxmas
today	tomorrow

friends

praise

fun

pray

good

weekend
Differential Language Analysis

Methods of Correlation Analysis:

- **Pearson Product-Moment Correlation**
  - Limitation: Doesn’t handle controls
Differential Language Analysis

Methods of Correlation Analysis:

- Pearson Product-Moment Correlation
  Limitation: Doesn’t handle controls

\[ r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}} \]
Differential Language Analysis

Methods of Correlation Analysis:

- Pearson Product-Moment Correlation
  Limitation: Doesn’t handle controls

- Standardized Multivariate Linear Regression
  Fit the model:

\[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i \]
Differential Language Analysis

Methods of Correlation Analysis:

- **Pearson Product-Moment Correlation**
  Limitation: Doesn’t handle controls

- **Standardized** Multivariate Linear Regression
  Fit the model:

\[
Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i
\]

Adjust all variables to have “mean center” and “unit variance”: 

\[
 r_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}
\]
Differential Language Analysis

Methods of Correlation Analysis:

- **Pearson Product-Moment Correlation**
  Limitation: Doesn’t handle controls

- **Standardized** Multivariate Linear Regression
  Fit the model:
  \[
  Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i
  \]

  Adjust all variables to have “mean center” and “unit variance”:
  \[
  z = \frac{x - \mu}{\sigma}
  \]

\[
\begin{align*}
\mu & = \text{Mean} \\
\sigma & = \text{Standard Deviation}
\end{align*}
\]
Differential Language Analysis

Methods of Correlation Analysis:

- **Pearson Product-Moment Correlation**
  Limitation: Doesn’t handle controls

- **Standardized Multivariate Linear Regression**
  Fit the model:
  $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i$$
  Option 1: Gradient Descent:
  $$J = \sum (y - \hat{y})^2$$ -- “Sum of Squares” Error
Differential Language Analysis

Methods of Correlation Analysis:

- Pearson Product-Moment Correlation
  Limitation: Doesn’t handle controls

- Standardized Multivariate Linear Regression
  Fit the model:
  Option 1: Gradient Descent:
  \[ J = \sum (y - \hat{y})^2 \quad \text{-- “Sum of Squares” Error} \]
  Option 2: Matrix model:
  \[ Y = X \beta + \epsilon \]
Differential Language Analysis

Methods of Correlation Analysis:

- Pearson Product-Moment Correlation
  Limitation: Doesn’t handle controls

- Standardized Multivariate Linear Regression
  Fit the model:
  \[ Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_m X_{m1} + \epsilon_i \]

  Option 1: Gradient Descent:
  \[ J = \sum (y - \hat{y})^2 \]  -- “Sum of Squares” Error

  Option 2: Matrix model:
  \[ \hat{Y} = X\beta + \epsilon \]

  Matrix Computation Solution:
  \[ \hat{\beta} = (X^T X)^{-1} X^T Y \]
Differential Language Analysis

Methods of Correlation Analysis:

- Pearson Product-Moment Correlation
  Limitation: Doesn’t handle controls

- Standardized Multivariate Linear Regression
  Fit the model:
  Option 1: Gradient Descent:
  \[ J = \sum (y - \hat{y})^2 \]  -- “Sum of Squares” Error
  Option 2: Matrix model:
  \[ Y = X\beta + \epsilon \]
  Matrix Computation Solution:
  \[ \hat{\beta} = (X^T X)^{-1} X^T Y \]
Differential Language Analysis

Methods of “Correlation” Analysis for binary outcomes:

- Logistic Regression over Standardized variables
- Odds Ratio

\[
\frac{\text{countA(“horrible”)}}{\text{NA}} - \frac{1}{\text{countA(“horrible”)}} = \frac{\text{countB(“horrible”)}}{\text{NA}} - \frac{1}{\text{countB(“horrible”)}}
\]

(Monroe et al., 2010; Jurafsky, 2017)
Differential Language Analysis

Methods of “Correlation” Analysis for binary outcomes:

- Logistic Regression over Standardized variables
- Odds Ratio

\[ \log \left( \frac{\text{countA}(“horrible”)}{\text{NA}} \right) - \log \left( \frac{\text{countB}(“horrible”)}{\text{NB}} \right) \]

(Monroe et al., 2010; Jurafsky, 2017)
Differential Language Analysis

Methods of “Correlation” Analysis for binary outcomes:

- Logistic Regression over Standardized variables
- Odds Ratio using Informative Dirichlet Prior

\[
\hat{\delta}_w^{(i-j)} = \log \left( \frac{y_i^w + \alpha_w}{n_i + \alpha_0 - (y_i^w + \alpha_w)} \right) - \log \left( \frac{y_i^j + \alpha_w}{n_i + \alpha_0 - (y_i^j + \alpha_w)} \right)
\]

(Monroe et al., 2010; Jurafsky, 2017)
Differential Language Analysis

Methods of “Correlation” Analysis for binary outcomes:

- Logistic Regression over Standardized variables
- Odds Ratio using **Informative Dirichlet Prior**

\[
\tilde{\delta}_w^{(i-j)} = \log \left( \frac{y^i_w + \alpha_w}{\alpha_0 + (y^i_w + \alpha_w)} \right) - \log \left( \frac{y^j_w + \alpha_w}{\alpha_0 + (y^j_w + \alpha_w)} \right)
\]

\[
\log \left( \frac{\text{countA("horrible")}}{\text{NA-countA("horrible")}} \right) - \log \left( \frac{\text{countB("horrible")}}{\text{NB-countB("horrible")}} \right)
\]

(Monroe et al., 2010; Jurafsky, 2017)
Differential Language Analysis

Methods of “Correlation” Analysis for binary outcomes:

- Logistic Regression over Standardized variables
- Odds Ratio using Informative Dirichlet Prior

Bayesian term for “smoothing”: accounts for uncertainty as a function of less events (i.e. words observed less) by integrating “prior” beliefs mathematically.

(Monroe et al., 2010; Jurafsky, 2017)
Differential Language Analysis

Methods of “Correlation” Analysis for binary outcomes:

- Logistic Regression over Standardized variables
- Odds Ratio using Informative Dirichlet Prior

\[ \delta_w^{(i-j)} = \log \left( \frac{y_w^i + \alpha_w}{n_i - \alpha_0 - (y_w^i + \alpha_w)} \right) \]

Bayesian term for “smoothing”: accounts for uncertainty as a function of less events (i.e. words observed less) by integrating “prior” beliefs mathematically.

“Informative”: the prior is based on past evidence. Here, the total frequency of the word.

(Monroe et al., 2010; Jurafsky, 2017)
Differential Language Analysis

Methods of “Correlation” Analysis for binary outcomes:

● Logistic Regression over Standardized variables

● Odds Ratio using Informative Dirichlet Prior

\[
\delta_{w}^{(i-j)} = \log\left(\frac{y_{w}^{i} + \alpha_{w}}{n^{i} + \alpha_{0} - (y_{w}^{i} + \alpha_{w})}\right) - \log\left(\frac{y_{w}^{j} + \alpha_{w}}{n^{j} + \alpha_{0} - (y_{w}^{j} + \alpha_{w})}\right)
\]

\[
\sigma^{2}\left(\delta_{w}^{(i-j)}\right) \approx \frac{1}{y_{w}^{i} + \alpha_{w}} + \frac{1}{y_{w}^{j} + \alpha_{w}}
\]

● Final statistic for a word: z-score of its log-odds-ratio:

\[
\frac{\delta_{w}^{(i-j)}}{\sqrt{\sigma^{2}\left(\delta_{w}^{(i-j)}\right)}}
\]

(Monroe et al., 2010; Jurafsky, 2017)
What this means for NLP:

1. Our data are inherently multi-level.
2. Often, there are “already-available” human attributes.
3. Our data and models are (human) biased.
Natural language is generated by people.

What this means for NLP:

Practical implication

1) More accurate models
2) Increased fairness in applications

Considering the people behind the language not only offers opportunities for improved accuracies but it could be fundamental to NLP’s role in our increasingly digital world.