Part-of-Speech Tagging for Historical English

Yi Yang and Jacob Eisenstein
Georgia Tech
Digital humanities research

- How does the portrayal of men and women differ in Shakespeare’s plays?
- What’s the language use patterns in North American slave narratives?

[Muralidharan and Hearst, 2011&2012]
Digital humanities research

How does the portrayal of men and women differ in Shakespeare’s plays?

What’s the language use patterns in North American slave narratives?

NLP can help!

[Muralidharan and Hearst, 2011&2012]
Digital humanities research

- How does the portrayal of men and women differ in Shakespeare’s plays?
- What’s the language use patterns in North American slave narratives?

- NLP can help!
- Only if NLP works for historical texts ...

[Muralidharan and Hearst, 2011&2012]
Hee said nobody had said anything agt mee.
Hee said nobody had said anything against me.

- Spelling variation

[Henry Oxinden, 1660]
Stanford POS Tagger

Stanford: NNP VBD NN VBD VBN NN NN NN

Hee said nobody had said anything agt mee.

- Spelling variation
Hee said nobody had said anything.

- Spelling variation
Transfer Loss for POS Tagging

Modern English

Error rate

[Rayson et al., 2007]
Transfer Loss for POS Tagging

Modern English: 3.0
Early Modern English: 18.0

[Rayson et al., 2007]
Approaches

- Spelling normalization
  - Map from historical spellings to contemporary forms.

- Rayson et al. (2007)
- Scheible et al. (2011)
- Bollmann (2011)
Approaches

- Spelling normalization
  - Map from historical spellings to contemporary forms.
- Domain adaptation (this work)
  - Build robust NLP systems with representation learning.

Rayson et al. (2007)
Scheible et al. (2011)
Bollmann (2011)

Yang & Eisenstein (2014)
Yang & Eisenstein (2015)
Original: Hee said nobody had said anything *agt mee*.

Normalized: Hee said nobody had said anything aged me.
Original: Hee said nobody had said anything *agt mee*.

Normalized: Hee said nobody had said anything aged me.

- Correct normalization

[VARD; Baron and Rayson, 2008]
Spelling Normalization

Original: Hee said nobody had said anything against me.

Normalized: Hee said nobody had said anything aged me.

- Correct normalization
- Incorrect normalization

[VARD; Baron and Rayson, 2008]
Spelling Normalization

Original: Hee said nobody had said anything against me.

Normalized: He said nobody had said anything aged me.

- Correct normalization
- Incorrect normalization
- False negative

[VARD; Baron and Rayson, 2008]
Spelling Normalization

Gold: PRP

Stanford: NNP VBD NN VBD VBN NN JJ PRP

Normalized: Hee said nobody had said anything aged me .

X X ✓

[VARD; Baron and Rayson, 2008]
Spelling Normalization

Gold: PRP IN
Stanford: N P VBD NN VBD VBN NN IN
Normalized: Hee said nobody had said anything aged me .

[VARD; Baron and Rayson, 2008]
Hee said nobody had said anything agt mee.
Hee said nobody had said anything agt mee.
Hee said nobody had said anything agt mee.
Hee said nobody had said anything agt mee.

<table>
<thead>
<tr>
<th>OOV Context</th>
<th>IV Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hee { said, was, came, told, ... }</td>
<td>He { I, We, ... } { said, was, came, told, ... }</td>
</tr>
</tbody>
</table>
Model
Hee said nobody had said anything *agt mee*.
Hee said nobody had said anything \textit{agt mee}.
Hee said nobody had said anything *agt mee* .

features

- CurrWord = hee
- NextWord = said
- Prefix1 = h
- Suffix1 = e
- ...

[FEMA; Yang and Eisenstein, 2015]
Hee said nobody had said anything agt mee.
Hee said nobody had said anything *agt mee*. 

**features**

- **CurrWord** = hee
- **NextWord** = said
- **Prefix1** = h
- **Suffix1** = e
- ...

[FEMA; Yang and Eisenstein, 2015]
Hee said nobody had said anything \texttt{agt mee}.
Feature Embeddings

\[ p(f_t | f_2) \propto \exp \left( u_2^T v_t \right) \]

features

\[ \begin{align*}
    \text{CurrWord} &= \text{hee} \\
    \text{NextWord} &= \text{said} \\
    \text{Prefix1} &= \text{h} \\
    \text{Suffix1} &= \text{e} \\
    \ldots
\end{align*} \]

Input embeddings

Output embeddings

\[ \begin{align*}
    v_1 &\quad \bullet \quad \bullet \quad \bullet \\
    v_2 &\quad \circ \quad \circ \quad \circ \\
    v_3 &\quad \bullet \quad \bullet \quad \bullet \\
    v_4 &\quad \bullet \quad \bullet \quad \bullet
\end{align*} \]

[FEMA; Yang and Eisenstein, 2015]
\[ p(f_t | f_2) \propto \exp (u_2^\top v_t) \]

\[ \ell = \sum_{t \neq 2} \log p(f_t | f_2) \]

\[ p(f_t | f_2) \propto \exp (u_2^\top v_t) \]

\[ \ell = \sum_{t \neq 2} \log p(f_t | f_2) \]

features

\{ 
  CurrWord = hee 
  NextWord = said 
  Prefix1 = h 
  Suffix1 = e 
  ... 
\}

[FEMA; Yang and Eisenstein, 2015]
Word Embeddings

words
- hee
- said
- nobody
- had
- ...

features
- currWord = hee
- nextWord = said
- prefix1 = h
- suffix1 = e
- ...

- Word embeddings
- Feature embeddings

[word2vec; Mikolov et al., 2013]
Word Embeddings

- **words**
  - hee
  - said
  - nobody
  - had
  - ...

- **features**
  - CurrWord = hee
  - NextWord = said
  - Prefix1 = h
  - Suffix1 = e
  - ...

- **Word embeddings**
- **Generic representations**
- **Feature embeddings**

[word2vec; Mikolov et al., 2013]
Word Embeddings

- **Words**
  - hee
  - said
  - nobody
  - had
  - ...

- **Features**
  - CurrWord = hee
  - NextWord = said
  - Prefix1 = h
  - Suffix1 = e
  - ...

- **Word embeddings**
- **Generic representations**
- **Feature embeddings**
- **Task-specific representations**

[word2vec; Mikolov et al., 2013]
Word Embeddings

- **Word embeddings**
  - Generic representations
  - Word co-occurrences

- **Feature embeddings**
  - Task-specific representations

### Words
- hee
- said
- nobody
- had
- ...

### Features
- CurrWord = hee
- NextWord = said
- Prefix1 = h
- Suffix1 = e
- ...

[word2vec; Mikolov et al., 2013]
Word Embeddings

words
- hee
- said
- nobody
- had
-...

features
- CurrWord = hee
- NextWord = said
- Prefix1 = h
- Suffix1 = e
-...

- Word embeddings
- Generic representations
- Word co-occurrences
- Feature embeddings
- Task-specific representations
- Feature co-occurrences

[word2vec; Mikolov et al., 2013]
Learning from Multiple Domains

- Previous work on unsupervised domain adaptation involves in two domains.

[FEMA; Yang and Eisenstein, 2015]
Learning from Multiple Domains

- Previous work on unsupervised domain adaptation involves in two domains.
- Unsupervised multi-domain adaptation

[FEMA; Yang and Eisenstein, 2015]
Previous work on unsupervised domain adaptation involves in two domains.

Unsupervised multi-domain adaptation
Hee said nobody had said anything agt mee.
### Multiple Feature Embeddings

<table>
<thead>
<tr>
<th>Domain Attributes:</th>
<th>Genre</th>
<th>Epoch</th>
</tr>
</thead>
</table>

Hee said nobody had said anything *agt mee*.  

[FEMA; Yang and Eisenstein, 2015]
Hee said nobody had said anything agt mee.
Multiple Feature Embeddings

Domain Attributes:

<table>
<thead>
<tr>
<th>Genre</th>
<th>Epoch</th>
</tr>
</thead>
<tbody>
<tr>
<td>letters</td>
<td>1600+</td>
</tr>
</tbody>
</table>

Hee said nobody had said anything *agt mee*.

features

- CurrWord = hee
- NextWord = said
- Prefix1 = h
- Suffix1 = e
- ...

[FEMA; Yang and Eisenstein, 2015]
Multiple Feature Embeddings

Domain Attributes:

- **Genre**: letters
- **Epoch**: 1600+

Hee said nobody had said anything *agt mee*.

Features:

1. CurrWord = hee
2. NextWord = said
   - Prefix1 = h
   - Suffix1 = e
   - ...

\[
\text{(shared)} + \text{(letters)} + \text{(1600+)}
\]

[FEMA; Yang and Eisenstein, 2015]
Hee said nobody had said anything agt mee.

Domain Attributes:
- Genre
- Epoch
  - letters
  - 1600+

Features:
- CurrWord = hee
- NextWord = said
- Prefix1 = h
- Suffix1 = e
- ...

\[
\begin{align*}
\text{features} &= \{\text{CurrWord = hee, NextWord = said, Prefix1 = h, Suffix1 = e, ...}\} \\
&= \begin{align*}
&= \begin{array}{c}
&\text{1} \quad \text{2} \quad \text{3} \quad \text{4} \\
&\text{features} = \{\text{currWord} = \text{hee, nextWord} = \text{said, prefix1} = \text{h, suffix1} = \text{e, ...}\} \\
&= \begin{array}{c}
&\text{(shared)} \quad \text{+} \quad \text{(letters)} \quad \text{+} \quad \text{(1600+)} \\
&\text{=}\end{array}
\end{align*}
\]

[FEMA; Yang and Eisenstein, 2015]
Hee said nobody had said anything *agt mee*.
Multiple Feature Embeddings

\[
\begin{align*}
\mathbf{u}_2 &= \mathbf{h}_2^{(\text{shared})} + \mathbf{h}_2^{(\text{letters})} + \mathbf{h}_2^{(1600+)} \\
\end{align*}
\]

Hee said nobody had said anything agt mee.

\[
\begin{align*}
\text{features} \left\{ \\
\text{CurrWord} &= \text{hee} \\
\text{NextWord} &= \text{said} \\
\text{Prefix1} &= \text{h} \\
\text{Suffix1} &= \text{e} \\
\cdots
\end{align*}
\]

[FEMA; Yang and Eisenstein, 2015]
Multiple Feature Embeddings

\[ p(f_t|f_2) \propto \exp(\mathbf{u}_2^\top \mathbf{v}_t) \]

\[ \mathbf{u}_2 = \mathbf{h}_2^{(\text{shared})} + \mathbf{h}_2^{(\text{letters})} + \mathbf{h}_2^{(1600+)} \]

“Hee” said nobody had said anything. 

features \[ \begin{align*}
\text{CurrWord} &= \text{hee} \\
\text{NextWord} &= \text{said} \\
\text{Prefix1} &= \text{h} \\
\text{Suffix1} &= \text{e} \\
\ldots
\end{align*} \]

[FEMA; Yang and Eisenstein, 2015]
Experiments
Penn Corpora of Historical English

Modern British English (MBE)

<table>
<thead>
<tr>
<th>Time Period</th>
<th># of Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>1840-1914</td>
<td>322,255</td>
</tr>
<tr>
<td>1770-1839</td>
<td>427,424</td>
</tr>
<tr>
<td>1700-1769</td>
<td>343,024</td>
</tr>
</tbody>
</table>

Early Modern English (EME)

<table>
<thead>
<tr>
<th>Time Period</th>
<th># of Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>1640-1710</td>
<td>614,315</td>
</tr>
<tr>
<td>1570-1639</td>
<td>706,587</td>
</tr>
<tr>
<td>1500-1569</td>
<td>640,255</td>
</tr>
</tbody>
</table>

[Kroch and Taylor, 2000; Kroch et al., 2004]
Tagset Mappings

- Penn Corpora of Historical English (PCHE) tagset: 83 tags
- Penn Treebank (PTB) tagset: 45 tags

[Moon and Baldridge, 2007]
Tagset Mappings

- Penn Corpora of Historical English (PCHE) tagset: 83 tags
- Penn Treebank (PTB) tagset: 45 tags

<table>
<thead>
<tr>
<th>PCHE</th>
<th>PTB</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJ</td>
<td>JJ</td>
</tr>
<tr>
<td>ADV</td>
<td>RB</td>
</tr>
<tr>
<td>ALSO</td>
<td>RB</td>
</tr>
<tr>
<td>VB</td>
<td>VB</td>
</tr>
<tr>
<td>VBI</td>
<td>VB</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Systems

- Support vector machine (SVM) tagger
  - Sixteen basic feature templates by Ratnaparkhi (1996)
Systems

- Support vector machine (SVM) tagger
  - Sixteen basic feature templates by Ratnaparkhi (1996)

- Representation learning methods
  - Structural correspondence learning (SCL)
  - Brown clustering
  - word2vec embeddings
  - Multiple feature embeddings (FEMA)

[Blitzer et al., 2006; Brown et al., 1992; Mikolov et al., 2013]
Temporal Adaptation

Modern British English (MBE)

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Train</th>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1840-1914</td>
<td>322,255</td>
<td>427,424</td>
<td>343,024</td>
</tr>
<tr>
<td>1770-1839</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1700-1769</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Early Modern English (EME)

<table>
<thead>
<tr>
<th>Time Period</th>
<th>Train</th>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1640-1710</td>
<td>614,315</td>
<td>706,587</td>
<td>640,255</td>
</tr>
<tr>
<td>1570-1639</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1500-1569</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

# of tokens
Results: Modern British English

Average error rate

Baseline: 4.6
Results: Modern British English

Average error rate

- Baseline: 4.6
- SCL: 4.3
- Brown: 4.2
- word2vec: 4.4
Results: Modern British English

Average error rate

Baseline: 4.6
SCL: 4.3
Brown: 4.2
word2vec: 4.4
FEMA: 3.7
(Our method): (-0.9)
Results: Early Modern English

Baseline: 9.4
Results: Early Modern English

Average error rate

Baseline: 9.4
SCL: 8.2
Brown: 8.0
word2vec: 8.3
Results: Early Modern English

Average error rate

Baseline: 9.4
SCL: 8.2
Brown: 8.0
word2vec: 8.3
FEMA: 6.6
(Our method): (- 2.8)
Adaptation from PTB

- Penn Treebank
  - Train: 969,905

- Modern British English
  - Test 1: 1,092,703

- Early Modern English
  - Test 2: 1,961,157

# of tokens

0  500,000  1,000,000  1,500,000  2,000,000
Adaptation from PTB

Standard evaluation scenario for English POS tagging.
Adaptation from PTB

Standard evaluation scenario for English POS tagging.

Insufficient data annotation for historical texts.

- Low resource languages
- Specific genres, styles, or epochs
Results: Modern British English

Error rate

Baseline

18.9
Results: Modern British English

Baseline: 18.9
SCL: 18.4
Brown: 18.4
word2vec: 18.3
Results: Modern British English

<table>
<thead>
<tr>
<th>Method</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>18.9</td>
</tr>
<tr>
<td>SCL</td>
<td>18.4</td>
</tr>
<tr>
<td>Brown</td>
<td>18.4</td>
</tr>
<tr>
<td>word2vec</td>
<td>18.3</td>
</tr>
<tr>
<td>FEMA</td>
<td>17.5 (-1.4)</td>
</tr>
</tbody>
</table>

(Our method)
Results: Early Modern English

Error rate

Baseline: 25.9
Results: Early Modern English

Error rate

<table>
<thead>
<tr>
<th>Method</th>
<th>Error Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>25.9</td>
</tr>
<tr>
<td>SCL</td>
<td>24.1</td>
</tr>
<tr>
<td>Brown</td>
<td>24.0</td>
</tr>
<tr>
<td>word2vec</td>
<td>24.2</td>
</tr>
</tbody>
</table>
Results: Early Modern English

Error rate

- Baseline: 25.9
- SCL: 24.1
- Brown: 24.0
- word2vec: 24.2
- FEMA: 22.1

(Our method) (-3.8)
Normalization vs. Representation Learning

Error rate

Baseline: 25.9
Representation learning: 22.1 (-3.8)
FEMA: -4.9
Normalization vs. Representation Learning

<table>
<thead>
<tr>
<th></th>
<th>Error Rate</th>
<th>Normalization vs. Representation Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>25.9</td>
<td></td>
</tr>
<tr>
<td>FEMA</td>
<td>22.1 (-3.8)</td>
<td></td>
</tr>
<tr>
<td>VARD</td>
<td>23.3 (-2.6)</td>
<td></td>
</tr>
</tbody>
</table>
Normalization vs. Representation Learning

<table>
<thead>
<tr>
<th>Error Rate</th>
<th>Baseline</th>
<th>Representation Learning</th>
<th>Spelling Normalization</th>
<th>Representation Learning + Normalization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25.9</td>
<td>22.1 (-3.8)</td>
<td>23.3 (-2.6)</td>
<td>21.0 (-4.9)</td>
</tr>
<tr>
<td></td>
<td>VARD</td>
<td>FEMA+ VARD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Error Analysis

- Annotation inconsistencies and tagset mismatches

<table>
<thead>
<tr>
<th>token</th>
<th>annotations in PCHE</th>
<th>annotations in PTB</th>
</tr>
</thead>
<tbody>
<tr>
<td>, (comma)</td>
<td>, (comma; 83.4%)</td>
<td>, (comma)</td>
</tr>
<tr>
<td>. (period; 16.6%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Error Analysis

- Annotation inconsistencies and tagset mismatches

<table>
<thead>
<tr>
<th>token</th>
<th>annotations in PCHE</th>
<th>annotations in PTB</th>
</tr>
</thead>
<tbody>
<tr>
<td>, (comma)</td>
<td>, (comma; 83.4%)</td>
<td>, (comma)</td>
</tr>
<tr>
<td>. (period)</td>
<td>, (comma; 12.3%)</td>
<td>. (period)</td>
</tr>
<tr>
<td></td>
<td>. (period; 16.6%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>. (period; 87.7%)</td>
<td></td>
</tr>
</tbody>
</table>
# Error Analysis

- Annotation inconsistencies and tagset mismatches

<table>
<thead>
<tr>
<th>token</th>
<th>annotations in PCHE</th>
<th>annotations in PTB</th>
</tr>
</thead>
<tbody>
<tr>
<td>, (comma)</td>
<td>, (comma; 83.4%) . (period; 16.6%)</td>
<td>, (comma)</td>
</tr>
<tr>
<td>. (period)</td>
<td>, (comma; 12.3%) . (period; 87.7%)</td>
<td>. (period)</td>
</tr>
<tr>
<td>to</td>
<td>TO (54.6%) IN (44.3%)</td>
<td>TO</td>
</tr>
</tbody>
</table>
## Error Analysis

- **Annotation inconsistencies and tagset mismatches**

<table>
<thead>
<tr>
<th>token</th>
<th>annotations in PCHE</th>
<th>annotations in PTB</th>
</tr>
</thead>
<tbody>
<tr>
<td>, (comma)</td>
<td>, (comma; 83.4%)</td>
<td>, (comma)</td>
</tr>
<tr>
<td>. (period)</td>
<td>. (comma; 12.3%) , (period; 87.7%)</td>
<td>. (period)</td>
</tr>
<tr>
<td>to</td>
<td>TO (54.6%) , IN (44.3%)</td>
<td>TO</td>
</tr>
<tr>
<td>all/any/every</td>
<td>JJ (quantifier)</td>
<td>DT</td>
</tr>
</tbody>
</table>
Conclusions
Conclusions

- Feature embeddings outperform word embeddings by exploiting task-specific information in feature templates.
Conclusions

- Feature embeddings outperform word embeddings by exploiting task-specific information in feature templates.

- Representation learning and spelling normalization are complementary for improving tagging performance.
Conclusions

- Feature embeddings outperform word embeddings by exploiting task-specific information in feature templates.
- Representation learning and spelling normalization are complementary for improving tagging performance.
- Tagset mismatches make it hard to evaluate modern POS taggers for historical English.