Natural Language Processing: Algorithms and Applications, Old and New

Noah Smith

Carnegie Mellon University $\rightarrow$ University of Washington

WSDM Winter School, January 31, 2015
Outline

I. Introduction to NLP
II. Algorithms for NLP
III. Example applications
Introduction to NLP
Why NLP?
What does it mean to “know” a language?
Levels of Linguistic Knowledge

- speech
- text
- phonetics
- orthography
- phonology
- morphology
- syntax
- semantics
- pragmatics
- discourse

"shallower"

"deeper"
ลูกศิษย์วัดกระทิงยังยื้อปิดถนนทางขึ้นไปนมัสการพระบาทเขาคิชฌกูฏ หวิดปะทะกับเจ้าถิ่นที่ออกมาเผชิญหน้าเพราะเดือดร้อนสัญจรไม่ได้ ผวจ.เร่งทุกฝ่ายเจรจา ก่อนที่ชื่อเสียงของจังหวัดจะเสียหายไปมากกว่านี้ พร้อมเสนอหยุดจัดงาน 15 วัน....
uygarlaştıramadıklarımızdanmışsınızcasına “(behaving) as if you are among those whom we could not civilize”
A ship-shipping ship, shipping shipping-ships.
(Syntactic knowledge required.)
Example: Part-of-Speech Tagging
(Gimpel et al., 2011; Owoputi et al., 2013)

ikr smh he asked fir yo last name

so he can add u on fb lololol
Example: Part-of-Speech Tagging

(Gimpel et al., 2011; Owoputi et al., 2013)

I know, right shake my head for your

ikr smh he asked fir yo last name

so he can add u on fb lololol

you Facebook laugh out loud
Example: Part-of-Speech Tagging

(Gimpel et al., 2011; Owoputi et al., 2013)

I know, right shake my head

ikr smh he asked fir yo last name

! G O V P D A N

interjection acronym pronoun verb prep. det. adj. noun

so he can add u on fb lololol

P O V V O P ∧!

preposition proper noun
Part II

Algorithms for NLP
A Starting Point: Categorizing Texts

Mosteller and Wallace (1963) automatically inferred the authors of the disputed *Federalist Papers*. Many other examples:

- News: politics vs. sports vs. business vs. technology ...
- Reviews of films, restaurants, products: positive vs. negative
- Email: spam vs. not
- What is the reading level of a piece of text?
- How influential will a scientific paper be?
- Will a piece of proposed legislation pass?
Categorizing Texts: A Standard Line of Attack

1. Human experts label some data.
2. Feed the data to a learning algorithm $L$ that constructs an automatic labeling function (classifier) $C$.
3. Apply that function to as much data as you want!
Categorizing Texts: Notation

- **Training examples**: \( \mathbf{x} = \langle x_1, x_2, \ldots, x_N \rangle \)
- **Their categorical labels**: \( \mathbf{y} = \langle y_1, y_2, \ldots, y_N \rangle \), each \( y_n \in \mathcal{Y} \)
- **A classifier** \( C \) seeks to map any \( x \) to the “correct” \( y \)
  \[ x \rightarrow C \rightarrow y \]
- **A learner** \( L \) infers \( C \) from \( x \) and \( y \)
  \[ x \rightarrow L \rightarrow C \]
  \[ y \rightarrow L \rightarrow C \]
Categorizing Texts: $C$

First, $\phi$ maps $\langle x, y \rangle$ into $\mathbb{R}^D$ (feature vector).

Then $C$ uses the vector to map into $\mathcal{Y}$.

▶ Linear models define:

$$C(x) = \arg\max_{y \in \mathcal{Y}} w^T \phi(x, y)$$

where $w \in \mathbb{R}^D$ is a vector of coefficients.

▶ Many non-linear options available as well (decision trees, neural networks, . . . ).
Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled, SECTION 1. COMPENSATION FOR WORK-RELATED INJURY. (a) AUTHORIZATION OF PAYMENT— The Secretary of the Treasury shall pay, out of money in the Treasury not otherwise appropriated, the sum of $46,726.30 to John M. Ragsdale as compensation for injuries sustained by John M. Ragsdale in June and July of 1952 while John M. Ragsdale was employed by the National Bureau of Standards. (b) SETTLEMENT OF CLAIMS— The payment made under subsection (a) shall be a full settlement of all claims by John M. Ragsdale against the United States for the injuries referred to in subsection (a). SEC. 2. LIMITATION ON AGENTS AND ATTORNEYS’ FEES. It shall be unlawful for an amount that exceeds 10 percent of the amount authorized by section 1 to be paid to or received by any agent or attorney in consideration of services rendered in connection with this Act. Any person who violates this section shall be guilty of an infraction and shall be subject to a fine in the amount provided in title 18, United States Code.
Example of a Linear Model

Probabilistic models define $p(Y = y \mid \phi(x, y) = f)$:

$$C(x) = \arg\max_{y \in \mathcal{Y}} p(Y = y \mid \phi(x, y) = f)$$

$$\quad = \arg\max_{y \in \mathcal{Y}} p(Y = y) \cdot p(\phi(x, y) = f \mid Y = y)$$

**Naïve Bayes** makes a strong assumption:

$$\ldots = \arg\max_{y \in \mathcal{Y}} p(Y = y) \prod_{d=1}^{D} p([\phi(x, y)]_d = f_d \mid Y = y)$$

$$\quad = \arg\max_{y \in \mathcal{Y}} \underbrace{\log p(Y = y)}_{w_{Y=y}} + \sum_{d=1}^{D} \underbrace{\log p([\phi(x, y)]_d = f_d \mid Y = y)}_{w_{Y=y,\phi_d=f_d}}$$
Naïve Bayes is a linear model and a probabilistic model.

- Another example that is both linear and probabilistic: (multinomial) logistic regression

- Not all linear models are probabilistic!

- Not all probabilistic models are linear!
$C(x) = \arg\max_{y \in \mathcal{Y}} \mathbf{w}^\top \phi(x, y)$
\begin{align*}
\langle x, y^3 \rangle & \quad \langle x, y^1 \rangle \\
\langle x, y^2 \rangle & \quad \langle x, y^4 \rangle 
\end{align*}
\[ \langle x, y^3 \rangle \]
\[ \langle x, y^1 \rangle \]
\[ \langle x, y^2 \rangle \]
\[ \langle x, y^4 \rangle \]

\[ f_1 \]
\[ f_2 \]
Categorizing Texts: \( L \)

Usually learning \( L \) involves choosing \( w \).

Often set up as an optimization problem:

\[
\hat{w} = \arg\min_{w: \Omega(w) \leq \tau} \frac{1}{N} \sum_{n=1}^{N} \text{loss}(x_n, y_n; w)
\]

\[
\text{Loss}(w)
\]

Example: classic multi-class support vector machine,

\[
\Omega(w) = \|w\|_2^2
\]

\[
\text{loss}(x, y; w) = -w^T \phi(x, y) + \max_{y' \in \mathcal{Y}} w^T \phi(x, y') + \begin{cases} 
0 & \text{if } y = y' \\
1 & \text{otherwise}
\end{cases}
\]

\[
\sum_{n=1}^{N} \text{loss}(x_n, y_n; w)
\]
Categorizing Texts: L

Usually learning $L$ involves choosing $\mathbf{w}$. Often set up as an optimization problem:

$$
\hat{\mathbf{w}} = \text{argmin}_{\mathbf{w}: \Omega(\mathbf{w}) \leq \tau} \frac{1}{N} \sum_{n=1}^{N} \text{loss}(x_n, y_n; \mathbf{w})
$$

Example: multinomial logistic regression with $\ell_2$ regularization,

$$
\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2
$$

$$
\text{loss}(x, y; \mathbf{w}) = -\mathbf{w}^\top \phi(x, y) + \log \sum_{y' \in \mathcal{Y}} \exp \mathbf{w}^\top \phi(x, y')
$$
What about $\Omega(w)$?

We usually constrain $w$ to fall in an $\ell_2$ ball:

$$\min_{w: \|w\|_2^2 \leq \tau} \text{Loss}(w) \equiv \min_w \text{Loss}(w) + c\|w\|_2^2$$
What about $\Omega(w)$?

We usually constrain $w$ to fall in an $\ell_2$ ball:

$$\min_{w: \|w\|_2^2 \leq \tau} \text{Loss}(w) \equiv \min_w \text{Loss}(w) + c\|w\|_2^2$$

Newer idea: use $\ell_1$ ball instead (lasso; Tibshirani, 1996).

$$\min_w \text{Loss}(w) + c \underbrace{\|w\|_1}_{D \sum_{d=1}^{D} |w_d|}$$
What about $\Omega(w)$?

We usually constrain $w$ to fall in an $\ell_2$ ball:

$$\min_{w: \|w\|_2^2 \leq \tau} Loss(w) \equiv \min_w Loss(w) + c \|w\|_2^2$$

Newer idea: use $\ell_1$ ball instead (lasso; Tibshirani, 1996).

$$\min_w Loss(w) + c \left(\sum_{d=1}^{D} |w_d|\right)$$

Even newer idea: use “$\ell_1$ of $\ell_2$” (group lasso; Yuan and Lin, 2006).
Visualizing the Lasso and Group Lasso

\[ \Omega(w) = |w_1| + |w_2| + |w_3| \]

See our tutorial from EACL (Martins et al., 2014).
Visualizing the Lasso and Group Lasso

\[ \Omega(w) = |w_1| + |w_2| + |w_3| \]

\[ \Omega(w) = \sqrt{w_1^2 + w_2^2} + |w_3| \]

See our tutorial from EACL (Martins et al., 2014).
In categorizing a document, only some sentences are relevant.

Groups: one group for every sentence in every training-set document.
  - All of the features (words) occurring in the sentence are in its group.

Special algorithms are required to learn with thousands/millions of overlapping groups.

See “Making the most of bag of words: sentence regularization with alternating direction method of multipliers,” Yogatama and Smith (2014).
### Text Categorization Example

#### IBM vs. Mac

<table>
<thead>
<tr>
<th>Sentence</th>
<th>Negative</th>
<th>Positive</th>
</tr>
</thead>
<tbody>
<tr>
<td>from: anonymized</td>
<td></td>
<td></td>
</tr>
<tr>
<td>subject: accelerating the macplus ... ;)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>lines: 15 we’re about ready to take a bold step into the 90s around here by accelerating our rather large collection of stock macplus computers.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>yes indeed, difficult to comprehend why anyone would want to accelerate a macplus, but that’s another story.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>suffice it to say, we can get accelerators easier than new machines.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hey, i don’t make the rules ...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>anyway, on to the purpose of this post: i’m looking for info on macplus accelerators.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>so far, i’ve found some lit on the novy accelerator and the micrmac multispeed accelerator.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>both look acceptable, but i would like to hear from anyone who has tried these.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>also, if someone would recommend another accelerator for the macplus, i’d like to hear about it.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>thanks for any time and effort you expend on this!</td>
<td></td>
<td></td>
</tr>
<tr>
<td>karl</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence</td>
<td>Negative</td>
<td>Positive</td>
</tr>
<tr>
<td>-------------------------------------------------------------------------</td>
<td>----------------</td>
<td>----------------</td>
</tr>
<tr>
<td>this film is one big joke: you have all the basics elements of romance (love at first sight, great passion, etc.) and gangster flicks (brutality, dagerous machinations, the mysterious don, etc.), but it is all done with the crudest humor.</td>
<td>(0.42)</td>
<td>(0.07)</td>
</tr>
<tr>
<td>it’s the kind of thing you either like viserally and immediately ”get” or you don’t.</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>that is a matter of taste and expectations.</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>i enjoyed it and it took me back to the mid80s, when nicolson and turner were in their primes.</td>
<td>(0.02)</td>
<td>(0.01)</td>
</tr>
<tr>
<td>the acting is very good, if a bit obviously tongue - in - cheek.</td>
<td>(0.01)</td>
<td>(0.01)</td>
</tr>
</tbody>
</table>
Categorizing Texts: Choosing a Learner

- Do you want posterior probabilities, or just labels?
- How interpretable does your model need to be?
- What background knowledge do you have about the data that can help?
- What methods do you understand well enough to explain to others?
- What methods will your team/boss/reader understand?
- What implementations are available?
- Cost, scalability, programming language, compatibility with your workflow, ...
- How well does it work (on held-out data)?
Categorizing Texts: Recipe

1. Obtain a pool of correctly categorized texts $\mathcal{D}$.
2. Define a feature function $\phi$ from hypothetically-labeled texts to feature vectors.
3. Select a parameterized function $C$ from feature vectors to categories.
4. Select $C$’s parameters $\mathbf{w}$ using training set $\langle \mathbf{x}, y \rangle \subset \mathcal{D}$ and learner $L$.
5. Predict labels using $C$ on a held-out sample from $\mathcal{D}$; estimate quality.
Instead of a finite, discrete set $\mathcal{Y}$, each input $x$ has its own $\mathcal{Y}_x$.

- E.g., $\mathcal{Y}_x$ is the set of POS sequences that could go with sentence $x$.

$|\mathcal{Y}_x|$ depends on $|x|$, often exponentially!

- Our 25-POS tagset gives as many as $25^{|x|}$ outputs.

$\mathcal{Y}_x$ can usually be defined as a set of interdependent categorization problems.

- Each word's POS depends on the POS tags of nearby words!
Decoding a Sequence

Abstract problem:

\[ x = \langle x[1], x[2], \ldots, x[L] \rangle \]
\[ \downarrow \]
\[ C \]
\[ \downarrow \]
\[ y = \langle y[1], y[2], \ldots, y[L] \rangle \]

Simple solution: categorize each \( x[\ell] \) separately.

But what if \( y[\ell] \) and \( y[\ell + 1] \) depend on each other?
Linear Models, Generalized to Sequences

\[ \hat{y} = \arg\max_{y \in Y_x} \mathbf{w}^\top \phi(x, y[1], \ldots, y[L]) \]
\[ \hat{y} = \arg\max_{y \in \mathcal{Y}_x} w^\top \phi(x, y[1], \ldots, y[L]) \]

\[ \hat{y} = \arg\max_{y \in \mathcal{Y}_x} w^\top \left( \sum_{\ell=2}^{L} \phi_{local}(x, \ell, y[\ell - 1], y[\ell]) \right) \]
HMMs are probabilistic; they define:

\[ p(x, y) = p(\text{stop} \mid y[L]) \prod_{\ell=1}^{L} p(x[\ell] \mid y[\ell]) \cdot p(y[\ell] \mid y[\ell-1]) \]

(\text{emission}) \cdot (\text{transition})

(where \( y[0] \) is defined to be a special start symbol).

Emission and transition counts can be treated as features, with coefficients equal to their log-probabilities.

\[ w^\top \phi_{\text{local}}(x, \ell, y[\ell-1], y[\ell]) = \log p(x[\ell] \mid y[\ell]) + \log p(y[\ell] \mid y[\ell-1]) \]

The probabilistic view is sometimes useful (we will see this later).
Finding the Best Sequence $y$: Intuition

If we knew $y[1 : L - 1]$, picking $y[L]$ would be easy:

$$\arg\max_{\lambda} w^T \phi_{local}(x, L, y[L - 1], \lambda) + w^T \left( \sum_{\ell=2}^{L-1} \phi_{local}(x, \ell, y[\ell - 1], y[\ell]) \right)$$
Finding the Best Sequence $y$: Notation

Let:

$$V[L-1, \lambda] = \max_{y[1:L-2]} \mathbf{w}^\top \left( \sum_{\ell=2}^{L-2} \phi_{local}(x, \ell, y[\ell-1], y[\ell]) \right)$$

$$+ \mathbf{w}^\top \phi_{local}(x, L-1, y[L-2], \lambda)$$

Our choice for $y[L]$ is then:

$$\operatorname{argmax}_{\lambda} \left( \max_{\lambda'} \mathbf{w}^\top \phi_{local}(x, L, \lambda', \lambda) + V[L-1, \lambda'] \right)$$
Finding the Best Sequence $y$: Notation

Let:

$$V[L-1, \lambda] = \max_{y[1:L-2]} \mathbf{w}^\top \left( \sum_{\ell=2}^{L-2} \phi_{local}(x, \ell, y[\ell-1], y[\ell]) \right)$$

$$+ \mathbf{w}^\top \phi_{local}(x, L-1, y[L-2], \lambda)$$

Note that:

$$V[L-1, \lambda] = \max_{\lambda'} V[L-2, \lambda'] + \mathbf{w}^\top \phi_{local}(x, L-1, \lambda', \lambda)$$

And more generally:

$$\forall \ell \in \{2, \ldots\}, \quad V[\ell, \lambda] = \max_{\lambda'} V[\ell-1, \lambda'] + \mathbf{w}^\top \phi_{local}(x, \ell, \lambda', \lambda)$$
Visualization

<table>
<thead>
<tr>
<th>N</th>
<th>O</th>
<th>Â</th>
<th>V</th>
<th>A</th>
<th>!</th>
</tr>
</thead>
<tbody>
<tr>
<td>ikr</td>
<td>smh</td>
<td>he</td>
<td>asked</td>
<td>fir</td>
<td>yo</td>
</tr>
</tbody>
</table>

...
Finding the Best Sequence \( y \): Algorithm

Input: \( x, \ w, \ \phi_{\text{local}}(\cdot, \cdot, \cdot, \cdot) \)

\[
\forall \lambda, \ V[1, \lambda] = 0. \\
\text{For } \ell \in \{2, \ldots, L\}:
\]

\[
\forall \lambda, \ V[\ell, \lambda] = \max_{\lambda'} V[\ell - 1, \lambda'] + \mathbf{w}^\top \phi_{\text{local}}(x, \ell, \lambda', \lambda)
\]

Store the “argmax” \( \lambda' \) as \( B[\ell, \lambda] \).

\[
y[L] = \arg\max_\lambda V[L, \lambda].
\]

\textit{Backtrack}. For \( \ell \in \{L - 1, \ldots, 1\} \):

\[
y[\ell] = B[\ell + 1, y[\ell + 1]]
\]

\[
\text{Return } \langle y[1], \ldots, y[L] \rangle.
\]
Visualizing and Analyzing Viterbi

<table>
<thead>
<tr>
<th>N</th>
<th>O</th>
<th>V</th>
<th>A</th>
<th>ikr</th>
<th>smh</th>
<th>he</th>
<th>asked</th>
<th>fir</th>
<th>yo</th>
<th>...</th>
</tr>
</thead>
</table>

...ikr smh he asked fir yo ...
1. What is sequence labeling useful for?
2. What are the features \( \phi \)?
3. How we learn the parameters \( \mathbf{w} \)?
ikr smh he asked fir yo last name
interjection acronym pronoun verb prep. det. adj. noun
!
P O V
preposition

so he can add u on fb lololol
POV
proper noun

V V O P ∧ !
Supersense Tagging

With Commander Chris Ferguson at the helm,

Atlantis touched down at Kennedy Space Center.
With Commander Chris Ferguson at the helm, Atlantis touched down at Kennedy Space Center.
Named Entity Recognition: Another Example

\[ x = \text{Britain sent warships across the English Channel Monday to rescue} \]
\[ y = \text{Britons stranded by Eyjafjallajökull's volcanic ash cloud}. \]
Named Entity Recognition: Features

<table>
<thead>
<tr>
<th></th>
<th>$\phi(x, y)$</th>
<th>$\phi(x, y')$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>bias:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>count of $i$ s.t. $y[i] = B$</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>count of $i$ s.t. $y[i] = I$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>count of $i$ s.t. $y[i] = O$</td>
<td>14</td>
<td>15</td>
</tr>
<tr>
<td><strong>lexical:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>count of $i$ s.t. $x[i] = Britain$ and $y[i] = B$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>count of $i$ s.t. $x[i] = Britain$ and $y[i] = I$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>count of $i$ s.t. $x[i] = Britain$ and $y[i] = O$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>downcased:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>count of $i$ s.t. $lc(x[i]) = britain$ and $y[i] = B$</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>count of $i$ s.t. $lc(x[i]) = britain$ and $y[i] = I$</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>count of $i$ s.t. $lc(x[i]) = britain$ and $y[i] = O$</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>count of $i$ s.t. $lc(x[i]) = sent$ and $y[i] = O$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>count of $i$ s.t. $lc(x[i]) = warships$ and $y[i] = O$</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
**Named Entity Recognition: Features**

<table>
<thead>
<tr>
<th>( \phi )</th>
<th>( \phi(x, y) )</th>
<th>( \phi(x, y') )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>shape:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>count of ( i ) s.t. ( \text{shape}(x[i]) = \text{B} ) and ( y[i] = \text{B} )</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>count of ( i ) s.t. ( \text{shape}(x[i]) = \text{I} ) and ( y[i] = \text{I} )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>count of ( i ) s.t. ( \text{shape}(x[i]) = \text{O} ) and ( y[i] = \text{O} )</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>prefix:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>count of ( i ) s.t. ( \text{pre}_1(x[i]) = \text{B} ) and ( y[i] = \text{B} )</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>count of ( i ) s.t. ( \text{pre}_1(x[i]) = \text{I} ) and ( y[i] = \text{I} )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>count of ( i ) s.t. ( \text{pre}_1(x[i]) = \text{O} )</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>count of ( i ) s.t. ( \text{pre}_1(x[i]) = \text{O} ) and ( y[i] = \text{O} )</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>count of ( i ) s.t. ( \text{shape}(\text{pre}_1(x[i])) = \text{A} ) and ( y[i] = \text{B} )</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>count of ( i ) s.t. ( \text{shape}(\text{pre}_1(x[i])) = \text{A} ) and ( y[i] = \text{I} )</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>count of ( i ) s.t. ( \text{shape}(\text{pre}_1(x[i])) = \text{A} ) and ( y[i] = \text{O} )</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>( \mathbb{I}{\text{shape}(\text{pre}_1(x[1])) = \text{A} \wedge y_1 = \text{B}} )</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( \mathbb{I}{\text{shape}(\text{pre}_1(x[1])) = \text{A} \wedge y[1] = \text{O}} )</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td><strong>gazetteer:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>count of ( i ) s.t. ( x[i] ) is in the gazetteer and ( y[i] = \text{B} )</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>count of ( i ) s.t. ( x[i] ) is in the gazetteer and ( y[i] = \text{I} )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>count of ( i ) s.t. ( x[i] ) is in the gazetteer and ( y[i] = \text{O} )</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>count of ( i ) s.t. ( x[i] = \text{sent} ) and ( y[i] = \text{O} )</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
he was willing to budge a little on

the price which means a lot to me.

See: “Discriminative lexical semantic segmentation with gaps: running the MWE gamut,” Schneider et al. (2014).
Multiword Expressions

he was willing to budge a little on

O O O O O B I O

the price which means a lot to me.

O O O B I I I I O

a little; means a lot to me

See: “Discriminative lexical semantic segmentation with gaps: running the MWE gamut,” Schneider et al. (2014).
Multiword Expressions

he was willing to budge a little on

O O O O B b i l

the price which means a lot to me.

O O O B I I I I O

a little; means a lot to me; budge ... on

See: “Discriminative lexical semantic segmentation with gaps: running the MWE gamut,” Schneider et al. (2014).
Mr President, Noah's ark was filled not with production factors, but with living creatures.

Noahs Arche war nicht voller Produktionsfaktoren, sondern Geschöpfe.

Dyer et al. (2013): a single “diagonal-ness” feature leads gains in translation (Bleu score).

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>model 4</th>
<th>fast_align</th>
<th>speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese → English</td>
<td>34.1</td>
<td>34.7</td>
<td>13×</td>
</tr>
<tr>
<td>French → English</td>
<td>27.4</td>
<td>27.7</td>
<td>10×</td>
</tr>
<tr>
<td>Arabic → English</td>
<td>54.5</td>
<td>55.7</td>
<td>10×</td>
</tr>
</tbody>
</table>
Other Sequence Decoding Problems

- Word transliteration
- Speech recognition
- Music transcription
- Gene identification

Add dimensions:
- Image segmentation
- Object recognition
- Optical character recognition
Recall that for categorization, we set up learning as **empirical risk minimization**:

\[
\hat{w} = \arg\min_{w : \Omega(w) \leq \tau} \frac{1}{N} \sum_{n=1}^{N} \text{loss}(x_n, y_n; w)
\]

Example loss:

\[
\text{loss}(x, y; w) = -w^\top \phi(x, y) + \max_{y' \in \mathcal{Y}_x} w^\top \phi(x, y')
\]
Structured Perceptron (Collins, 2002)

Input: $x, y, T$, step size sequence $\langle \alpha_1, \ldots, \alpha_T \rangle$

1. $w = 0$
2. For $t \in \{1, \ldots, T\}$:
   1. Draw $n$ uniformly at random from $\{1, \ldots, N\}$.
   2. Decode $x_n$:
      $$\hat{y} = \arg\max_{y \in Y_{x_n}} w^\top \phi(x_n, y)$$
   3. If $\hat{y} \neq y_n$, update parameters:
      $$w = w + \alpha_t (\phi(x_n, y_n) - \phi(x_n, \hat{y}))$$
3. Return $w$
Variations on the Structured Perceptron

Change loss:

- **Conditional random fields**: use “softmax” instead of max in loss; generalizes logistic regression
- **Max-margin Markov networks**: use cost-augmented max in loss; generalizes support vector machine

Incorporate regularization $\Omega(w)$, as previously discussed.

Change the optimization algorithm:

- Automatic step-size scaling (e.g., MIRA, Adagrad)
- Batch and “mini-batch” updating
- Averaging and voting
Structured Prediction: Lines of Attack

1. Transform into a sequence of classification problems.
2. Transform into a sequence labeling problem and use a variant of the Viterbi algorithm.
3. Design a representation, prediction algorithm, and learning algorithm for your particular problem.
Beyond Sequences

- Can all linguistic structure be captured with sequence labeling?
- Some representations are more elegantly handled using other kinds of output structures.
  - Syntax: trees
  - Semantics: graphs
- Dynamic programming and other combinatorial algorithms are central.
  - Always useful: features $\phi$ that decompose into local parts
I ❤️ the Biebs & want to have his babies! –>

LA Times:

Teen Pop Star Heartthrob is All the Rage on Social Media

... #belieber

See: “A dependency parser for tweets,” Kong et al. (2014)
The boy wants to visit New York City.

See: “A discriminative graph-based parser for the Abstract Meaning Representation,” Flanigan et al. (2014)
Part III

Example Applications
Machine Translation

302  云南芫爆松茸
Sautéed trichroma matsutake with coriander and
细嫩，香味浓溢

303  白油爆鸡枞
Stir-fried wikipedia
肉质细嫩，洁白如玉，或炒或蒸，串汤作菜，清香四

304  香油鸡枞蒸水蛋
Steam eggs with wikipedia

305  寸金蒜片油鸡枞
Translation from Analytic to Synthetic Languages

How to generate well-formed words in a morphologically rich target language?

Useful tool: morphological lexicon

\[ y_\sigma = \text{пытаться} \]
\[ y_\mu = \{\text{Verb, MAIN, IND, PAST, SING, FEM, MEDIAL, PERF}\} \]

“Translating into morphologically rich languages with synthetic phrases,” Chahuneau et al. (2013)
Contemporary translation is performed by mapping source-language "phrases" to target-language "phrases."

A phrase is a sequence of one or more words.

In addition, let a phrase be a sequence of one or more *stems*.

Our approach automatically inflects stems in context, and lets these *synthetic* phrases compete with traditional ones.
Predicting Inflection in Multilingual Context

\[ y_\sigma = \text{пытаться} \]
\[ y_\mu = \{\text{Verb, MAIN, IND, PAST, SING, FEM, MEDIAL, PERF}\} \]

она пыталась пересечь пути на ее велосипеде, 

\[ \phi(x, y_\mu) = \langle \phi_{\text{source}}(x) \otimes \phi_{\text{target}}(y_\mu), \phi_{\text{target}}(y_\mu) \otimes \phi_{\text{target}}(y_\mu) \rangle \]
## Translation Results (out of English)

<table>
<thead>
<tr>
<th></th>
<th>→ Russian</th>
<th>→ Hebrew</th>
<th>→ Swahili</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>14.7±0.1</td>
<td>15.8±0.3</td>
<td>18.3±0.1</td>
</tr>
<tr>
<td>+Class LM</td>
<td>15.7±0.1</td>
<td>16.8±0.4</td>
<td>18.7±0.2</td>
</tr>
<tr>
<td>+Synthetic</td>
<td>16.2±0.1</td>
<td>17.6±0.1</td>
<td>19.0±0.1</td>
</tr>
</tbody>
</table>

Translation quality (Bleu score; higher is better), averaged across three runs.
Something Completely Different
Measuring Ideological Proportions

“Well, I think you hit a reset button for the fall campaign. Everything changes. It’s almost like an Etch-A-Sketch. You can kind of shake it up and restart all over again.”

—Eric Fehrnstrom, Mitt Romney’s spokesman, 2012
Measuring Ideological Proportions

“Well, I think you hit a reset button for the fall campaign. Everything changes. It’s almost like an Etch-A-Sketch. You can kind of shake it up and restart all over again.”

—Eric Fehrnstrom, Mitt Romney’s spokesman, 2012
Hypothesis: primary candidates “move to the center” before a general election.

- In primary elections, voters tend to be ideologically concentrated.
- In general elections, voters are now more widely dispersed across the ideological spectrum.

Do Obama, McCain, and Romney use more “extreme” ideological rhetoric in the primaries than the general election?

Can we measure candidates’ ideological positions from the text of their speeches at different times?

See: “Measuring ideological proportions in political speeches,” Sim et al. (2013).
Operationalizing “Ideology”

- Left
  - Religious Left
  - Progressive
  - Far Left
- Center
  - Center Left
- Right
  - Religious Right
  - Libertarian
  - Populist
  - Far Right
Instead of putting more limits on your earnings and your options, we need to place clear and firm limits on government spending. As a start, I will lower federal spending to 20 percent of GDP within four years’ time – down from the 24.3 percent today.
The President’s plan assumes an endless expansion of government, with costs rising and rising with the spread of Obamacare. I will halt the expansion of government, and repeal Obamacare.
Working together, we can save Social Security without making any changes in the system for people in or nearing retirement. We have two basic options for future retirees: a tax increase for high-income retirees, or a decrease in the benefit growth rate for high-income retirees. I favor the second option; it protects everyone in the system and it avoids higher taxes that will drag down the economy.
I have proposed a Medicare plan that improves the program, keeps it solvent, and slows the rate of growth in health care costs.
—Excerpt from speech by Romney on 5/25/12 in Des Moines, IA
Instead of putting more limits on your earnings and your options, we need to place clear and firm limits on government spending. As a start, I will lower federal spending to 20 percent of GDP within four years’ time – down from the 24.3 percent today.

The President’s plan assumes an endless expansion of government, with costs rising and rising with the spread of Obamacare. I will halt the expansion of government, and repeal Obamacare.

Working together, we can save Social Security without making any changes in the system for people in or nearing retirement. We have two basic options for future retirees: a tax increase for high-income retirees, or a decrease in the benefit growth rate for high-income retirees. I favor the second option; it protects everyone in the system and it avoids higher taxes that will drag down the economy.

I have proposed a Medicare plan that improves the program, keeps it solvent, and slows the rate of growth in health care costs.

—Excerpt from speech by Romney on 5/25/12 in Des Moines, IA
Cue-Lag Representation of a Speech

government spending 8 federal spending 47 repeal Obamacare 7

Social Security 24 tax increase 13 growth rate 21 higher taxes 29

health care costs
Line of Attack

1. Build a “dictionary” of cues.
2. Infer ideological proportions from the cue-lag representation of speeches.
Ideological Books Corpus

Center Left

Religious Left

Center

Religious Right

Progressive

Center Right

Libertarian

Far Left

Populist

Far Right
Ideological Books Corpus

- Center Left
- Religious Left
- Progressive
- Far Left
- Center
- Religious Right
- Center Right
- Libertarian
- Populist
- Far Right

Books:
- The Conservative Soul
- Eradicate Blowing Out God's America
- Government Bullies
- No, They Can't
- Lies the Government Told You
- On the Middle Class
- America by Heart
- Ameritopia
- Glenn Beck
- Sarah Palin
- Cowards
### Example Cues

<table>
<thead>
<tr>
<th>Political Ideology</th>
<th>Representative Authors/Books</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Center-Right</strong></td>
<td>D. Frum, M. McCain, C. T. Whitman (1,450)</td>
<td>governor bush; class voter; health care; republican president; george bush; state police; move forward; miss america; middle eastern; water buffalo; fellow citizens; sam’s club; american life; working class; general election; culture war; status quo; human dignity; same-sex marriage</td>
</tr>
<tr>
<td><strong>Libertarian</strong></td>
<td>Rand Paul, John Stossel, <em>Reason</em> (2,268)</td>
<td>medical marijuana; raw milk; rand paul; economic freedom; health care; government intervention; market economies; commerce clause; military spending; government agency; due process; drug war; minimum wage; federal law; ron paul; private property</td>
</tr>
<tr>
<td><strong>Religious Right</strong></td>
<td>(960)</td>
<td>daily saint; holy spirit; matthew [c/v]; john [c/v]; jim wallis; modern liberals; individual liberty; god’s word; jesus christ; elementary school; natural law; limited government; emerging church; private property; planned parenthood; christian nation; christian faith</td>
</tr>
</tbody>
</table>

Cue-Lag Ideological Proportions Model

Libertarian (R) → Libertarian (R) → Right → Progressive (L)

government spending

federal spending

repeal Obamacare

Social Security

Each speech is modeled as a sequence:
- ideologies are labels ($y$)
- cue terms are observed ($x$)
HMM “with a Twist”

Right  
repeal Obamacare

  
Progressive (L)  
Social Security
HMM “with a Twist”

\[ \mathbf{w}^T \phi_{local}(x, \ell, \text{Right}, \text{Prog.}) = \log p(\text{Right} \sim \text{Prog.}) + \ldots \]
HMM “with a Twist”

Also considers probability of restarting the walk through a “noisy-OR” model.
Learning and Inference

We do not have labeled examples $\langle x, y \rangle$ to learn from!

Instead, labels are “hidden.”

We sample from the posterior over labels, $p(y \mid x)$.

This is sometimes called approximate Bayesian inference.
Measuring Ideological Proportions in Speeches

- Campaign speeches from 21 candidates, separated into primary and general elections in 2008 and 2012.

- Run model on each candidate separately with
  - independent transition parameters for each epoch, but
  - shared emission parameters for a candidate.
Mitt Romney

Primaries 2012
General 2012
Religious (L)
Center-Right
Libertarian (R)
Religious (R)
Far Left
Progressive (L)
Left
Center-Left
Right
Populist (R)
Far Right
Barack Obama

Primaries 2008
General 2008
Far Left
Religious (L)
Left
Center-Left
Center-Right
Libertarian (R)
Populist (R)
Religious (R)
Progressive (L)
Center
Right
Far Right

Primaries 2008
Far Left
Religious (L)
Libertarian (R)
Center-Right
Populist (R)
Right
Far Right

General 2008
Center
Progressive (L)
Left
Religious (R)
Center-Left
Far Left

Primaries 2008
General 2008
John McCain

Primaries 2008

General 2008

Far Left

Religious (L)

Center-Left

Center-Right

Libertarian (R)

Religious (R)

Progressive (L)

Left

Center

Right

Populist (R)

Far Right

Progressive (L)

Left

Center

Right

Populist (R)

Far Right

Primaries 2008

General 2008
Objective Evaluation?

Pre-registered hypothesis
A statement by a domain expert about his/her *expectations* of the model’s output.
## Preregistered Hypotheses

<table>
<thead>
<tr>
<th>Hypotheses</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sanity checks (strong):</strong></td>
</tr>
<tr>
<td>S1. Republican primary candidates should tend to draw more from <strong>Right</strong> than from <strong>Left</strong>.</td>
</tr>
<tr>
<td>S2. Democratic primary candidates should tend to draw more from <strong>Left</strong> than from <strong>Right</strong>.</td>
</tr>
<tr>
<td>S3. In general elections, Democrats should draw more from the <strong>Left</strong> than the Republicans and vice versa for the <strong>Right</strong>.</td>
</tr>
<tr>
<td><strong>Primary hypotheses (strong):</strong></td>
</tr>
<tr>
<td>P1. Romney, McCain and other Republicans should almost never draw from <strong>Far Left</strong>, and extremely rarely from <strong>Progressive</strong>.</td>
</tr>
<tr>
<td>P2. Romney should draw more heavily from the <strong>Right</strong> than Obama in both stages of the 2012 campaign.</td>
</tr>
<tr>
<td><strong>Primary hypotheses (moderate):</strong></td>
</tr>
<tr>
<td>P3. Romney should draw more heavily on words from the <strong>Libertarian</strong>, <strong>Populist</strong>, <strong>Religious Right</strong>, and <strong>Far Right</strong> in the primary compared to the general election. In the general election, Romney should draw more heavily on <strong>Center</strong>, <strong>Center-Right</strong> and <strong>Left</strong> vocabularies.</td>
</tr>
</tbody>
</table>
Baselines

Compare against “simplified” versions of the model:

- HMM: traditional HMM without ideological tree structure
- NoRes: weaker assumptions (never restart)
- Mix: stronger assumptions (always restart)
## Results

<table>
<thead>
<tr>
<th></th>
<th>CLIP</th>
<th>HMM</th>
<th>Mix</th>
<th>NoRes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sanity checks</td>
<td>20/21</td>
<td>19/22</td>
<td>21/22</td>
<td>17/22</td>
</tr>
<tr>
<td>Strong hypotheses</td>
<td>31/34</td>
<td>23/33</td>
<td>28/34</td>
<td>30/34</td>
</tr>
<tr>
<td>Moderate hypotheses</td>
<td>14/17</td>
<td>14/17</td>
<td>12/17</td>
<td>11/17</td>
</tr>
<tr>
<td>Total</td>
<td>65/72</td>
<td>56/72</td>
<td>61/73</td>
<td>58/73</td>
</tr>
</tbody>
</table>
Summary

I Introduction to NLP

II Algorithms for NLP
   ▶ Categorizing Texts
      ▶ Sparsity and group sparsity
   ▶ Decoding Sequences
      ▶ Viterbi
      ▶ Structured perceptron
   ▶ Many examples of tasks

III Example Applications
   ▶ A translation problem
   ▶ A political science problem
Some Current Research Directions in NLP

- Representations for semantics
  - Distributed
  - Denotational
  - Non-propositional
  - Hybrids of all of the above
  - Broad-coverage as well as domain-specific

- Alternatives to annotating data:
  - Constraints and bias
  - Regularization and priors
  - Semisupervised learning
  - Feature/representation learning \(\approx\) unsupervised discovery

- Multilinguality

- Approximate inference algorithms for learning and decoding
Thank you!


References II


