Natural Language Processing: Algorithms and Applications, Old and New

Noah Smith

Carnegie Mellon University $\xrightarrow{2015}$ University of Washington

WSDM Winter School, January 31, 2015

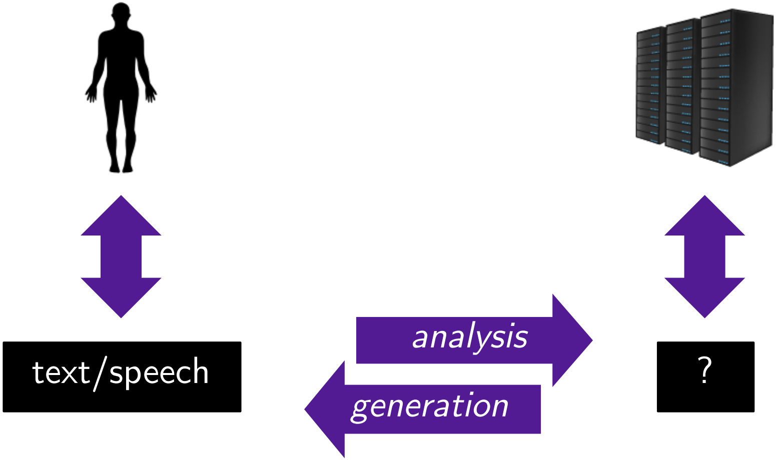
Outline

I. Introduction to NLP
II. Algorithms for NLP
III. Example applications

Part I

Introduction to NLP

Why NLP?


text/speech
analysis
generation
?

What does it mean to “know” a language?

Levels of Linguistic Knowledge

speech

text

phonetics

orthography

phonology

morphology

"shallower"

syntax

semantics

"deeper"

pragmatics

discourse

Orthographic Knowledge Required

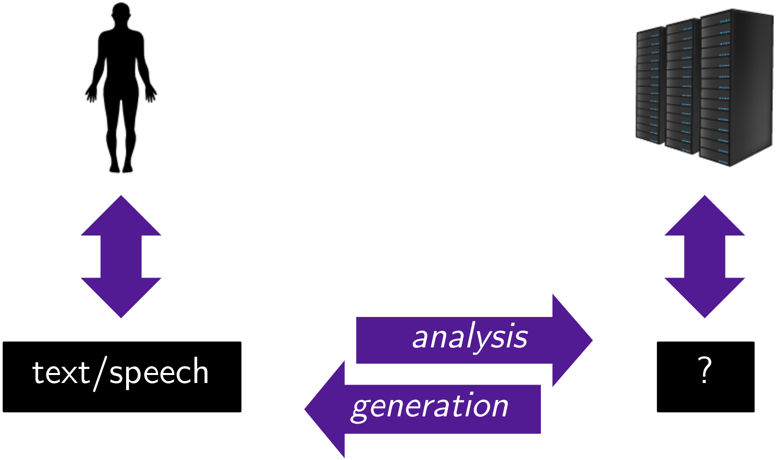
ลูกศิษย์วัดกระทิงยังยื้อปิดถนนทางขึ้นไปนมัสการพระบาทเขาคิชฌกูฏ หวิดปะทะกับเจ้าถิ่นที่ออกมาเผชิญหน้าเพราะเดือดร้อนสัญจรไม่ได้ ผวจ.เร่งทุกฝ่ายเจรจาก่อนที่ชื่อเสียงของจังหวัดจะเสียหายไปมากกว่านี้ พร้อมเสนอหยุดจัดงาน 15 วัน....

Morphological Knowledge Required

uygarlaştıramadıklarımızdanmışsınızcasına "(behaving) as if you are among those whom we could not civilize"

A ship-shipping ship, shipping shipping-ships.

(Syntactic knowledge required.)


text/speech
analysis
generation
?

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

				you		Facebook	laugh out loud
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
!	G	O	V	P	D	A	N
interjection	acronym	pronoun	verb	prep.	det.	adj.	noun

				you		Facebook	laugh out loud
so	he	can	add	u	on	fb	lololol
P	O	V	V	O	P	$\wedge$	!
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: postive 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

$$x \rightarrow \boxed{C} \rightarrow y$$

- A **learner** L infers C from **x** and **y**

$$\begin{array}{ll} \mathbf{x} & \rightarrow \\ \mathbf{y} & \rightarrow \end{array} \boxed{L} \rightarrow C$$

Categorizing Texts: C

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

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

- Linear models define:

$$C(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} \mathbf{w}^\top \phi(x, y)$$

  where $\mathbf{w} \in \mathbb{R}^D$ is a vector of coefficients.
- Many *non-linear* options available as well (decision trees, neural networks, ...).

Categorizing Texts

Example from Yano et al. (2012)

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 $p(Y = y \mid \phi(x, y) = \mathbf{f})$:

$$\begin{aligned} \color{blue}{C}(x) &= \operatorname*{argmax}_{y \in \mathcal{Y}} p(Y = y \mid \phi(x, y) = \mathbf{f}) \\ &= \operatorname*{argmax}_{y \in \mathcal{Y}} \frac{p(Y = y) \cdot p(\phi(x, y) = \mathbf{f} \mid Y = y)}{\color{gray}{p(\phi(x, y) = \mathbf{f})}} \end{aligned}$$

Naïve Bayes makes a strong assumption:

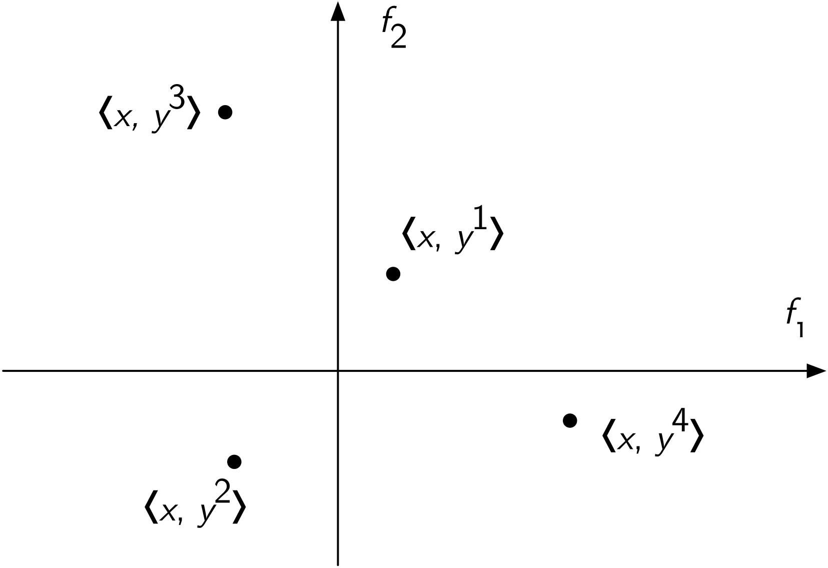
$$\begin{aligned} \ldots &= \operatorname*{argmax}_{y \in \mathcal{Y}} p(Y = y) \prod_{d=1}^{D} p([\phi(x, y)]_d = f_d \mid Y = y) \\ &= \operatorname*{argmax}_{y \in \mathcal{Y}} \underbrace{\log p(Y = y)}_{w_{Y=y}} + \sum_{d=1}^{D} \underbrace{\log p([\phi(x, y)]_d = f_d \mid Y = y)}_{w_{Y=y, \phi_d = f_d}} \end{aligned}$$

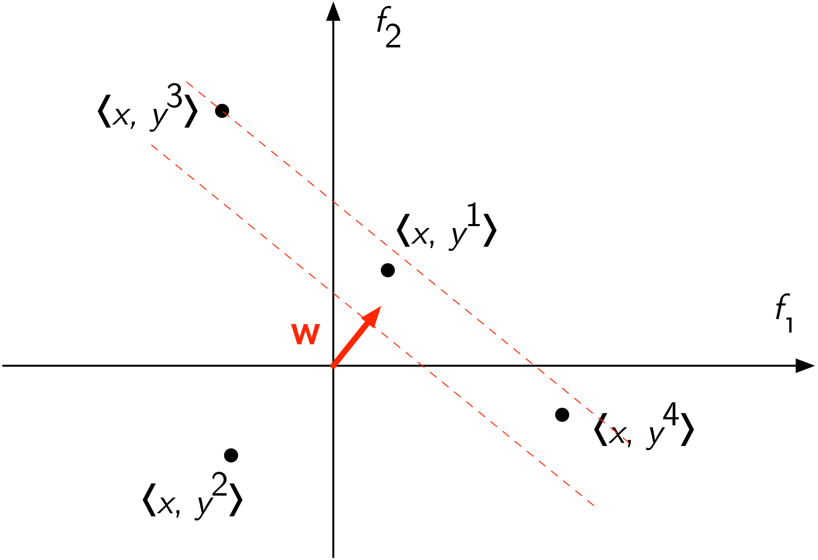
Note

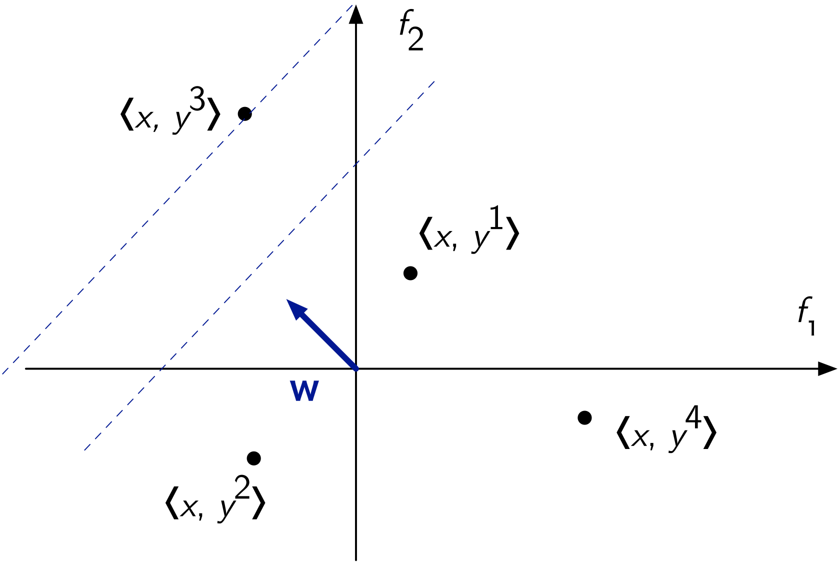
- Naïve Bayes is a linear model and a probabilistic model.
  - Another example that is both linear and probabilistic: **(multinomial) logistic regression**
- Not all linear models are probabilistic!
- Not all probabilistic models are linear!

C as Linear Model

$$C(x) = \operatorname*{argmax}_{y \in \mathcal{Y}} \mathbf{w}^\top \phi(x, y)$$


f_2
$\langle x, y^3 \rangle$
$\langle x, y^1 \rangle$
f_1
$\langle x, y^4 \rangle$
$\langle x, y^2 \rangle$


f_2
$\langle x, y^3 \rangle$
$\langle x, y^1 \rangle$
f_1
$\mathbf{w}$
$\langle x, y^4 \rangle$
$\langle x, y^2 \rangle$



f_2
$\langle x, y^3 \rangle$
$\langle x, y^1 \rangle$
f_1
$\mathbf{w}$
$\langle x, y^4 \rangle$
$\langle x, y^2 \rangle$

Categorizing Texts: L

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

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

Example: classic multi-class support vector machine,

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

$$loss(x, y; \mathbf{w}) = -\mathbf{w}^\top \phi(x, y) + \max_{y' \in \mathcal{Y}} \mathbf{w}^\top \phi(x, y') + \begin{cases} 0 & \text{if } y = y' \\ 1 & \text{otherwise} \end{cases}$$

Categorizing Texts: L

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

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

Example: multinomial logistic regression with ℓ_2 regularization,

$$\begin{aligned} \Omega(\mathbf{w}) &= \|\mathbf{w}\|_2^2 \\ loss(x, y; \mathbf{w}) &= -\mathbf{w}^\top \phi(x, y) + \log \sum_{y' \in \mathcal{Y}} \exp \mathbf{w}^\top \phi(x, y') \end{aligned}$$

What about $\Omega(\mathbf{w})$?

We usually constrain $\mathbf{w}$ to fall in an ℓ_2 ball:

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

What about $\Omega(\mathbf{w})$?

We usually constrain $\mathbf{w}$ to fall in an ℓ_2 ball:

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

Newer idea: use ℓ_1 ball instead (lasso; Tibshirani, 1996).

$$\min_{\mathbf{w}} Loss(\mathbf{w}) + c \underbrace{\|\mathbf{w}\|_1}_{\sum_{d=1}^{D} |w_d|}$$

What about $\Omega(\mathbf{w})$?

We usually constrain $\mathbf{w}$ to fall in an ℓ_2 ball:

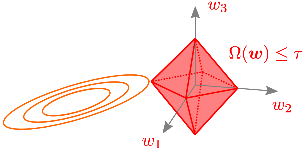
$$\min_{\mathbf{w}:\|\mathbf{w}\|_2^2 \leq \tau} Loss(\mathbf{w}) \quad \equiv \quad \min_{\mathbf{w}} Loss(\mathbf{w}) + c\|\mathbf{w}\|_2^2$$

Newer idea: use ℓ_1 ball instead (lasso; Tibshirani, 1996).

$$\min_{\mathbf{w}} Loss(\mathbf{w}) + c \underbrace{\|\mathbf{w}\|_1}_{\sum_{d=1}^{D} |w_d|}$$

Even newer idea: use "ℓ_1 of ℓ_2" (group lasso; Yuan and Lin, 2006).

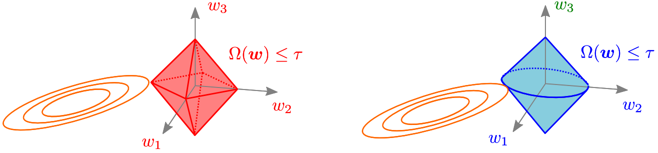
Visualizing the Lasso and Group Lasso



$$\boxed{\Omega(\boldsymbol{w}) = |w_1| + |w_2| + |w_3|}$$

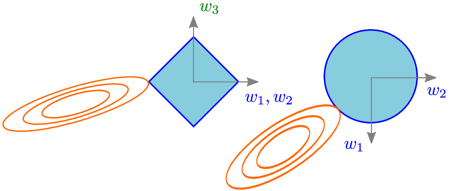
See our tutorial from EACL (Martins et al., 2014).

Visualizing the Lasso and Group Lasso



$$\boxed{\Omega(\boldsymbol{w}) = |w_1| + |w_2| + |w_3|}$$

$$\boxed{\Omega(\boldsymbol{w}) = \sqrt{w_1^2 + w_2^2} + |w_3|}$$



See our tutorial from EACL (Martins et al., 2014).

Using Data to Create Group Lasso's Groups

(Yogatama and Smith, 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

Sentence	Negative	Positive
from : *anonymized*		
subject : accelerating the macplus ... ;)		(0.05)
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 .		(0.07) (0.03) (0.02) (0.02)
yes indeed , difficult to comprehend why anyone would want to accelerate a macplus, but that's another story .		(0.06) (0.02) (0.02) (0.04)
suffuce it to say , we can get accelerators easier than new machines .		(0.01)
hey , i don ' t make the rules ...		(0.01)
anyway , on to the purpose of this post: i ' m looking for info on macplus acelerators .		(0.04) (0.01)
so far , i ' ve found some lit on the novy accelerator and the micrmac multispeed accelartor .		(0.02) (0.02) (0.02) (0.04)
both look acceptable , but i would like to hear from anyone who has tried these .	(–0.01)	
also , if someone would recommend another accelerator for the macplus , i ' d like to hear about it .		(0.06) (0.03) (0.02) (0.06)
thanks for any time and effort you expend on this !	(–0.01) (–0.01) (–0.01)	
karl		

Sentiment Analysis

Amazon DVDs (Blitzer et al., 2007)

Sentence	Negative Positive
this film is one big joke : you have all the basics elements	(0.42)
of romance (love at first sight , great passion , etc .) and gangster flicks	(0.22)
(brutality , dagerous machinations , the mysterious don , etc.) ,	(0.07)
but it is all done with the crudest humor .	(0.48)
it ' s the kind of thing you either like viserally and	(0.01)
immediately " get " or you don ' t .	(0.01)
that is a matter of taste and expectations .	(0.01)
i enjoyed it and it took me back to the mid80s ,	(0.02)
when nicolson and turner were in their primes .	(0.01)
the acting is very good , if a bit obviously tongue - in - cheek .	(0.01)

Categorizing Texts: Choosing a Learner L

- 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 ϕ 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 $\langle \mathbf{x}, \mathbf{y} \rangle \subset \mathcal{D}$ and learner L.
5. Predict labels using C on a held-out sample from $\mathcal{D}$; estimate quality.

From Categorization to Structured Prediction

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:

$$
\begin{array}{ccc}
x & = & \langle x[1], x[2], \ldots, x[L] \rangle \\
\downarrow & & \\
\boxed{C} & & \\
\downarrow & & \\
y & = & \langle y[1], y[2], \ldots, y[L] \rangle
\end{array}
$$

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

$$\hat{y} = \operatorname*{argmax}_{y \in \mathcal{Y}_x} \mathbf{w}^\top \phi(x, y[1], \ldots, y[L])$$

Linear Models, Generalized to Sequences

$$\hat{y} = \operatorname*{argmax}_{y \in \mathcal{Y}_x} \mathbf{w}^\top \phi(x, y[1], \ldots, y[L])$$

$$\hat{y} = \operatorname*{argmax}_{y \in \mathcal{Y}_x} \mathbf{w}^\top \left(\sum_{\ell=2}^{L} \phi_{local}(x, \ell, y[\ell-1], y[\ell]) \right)$$

Special Case: Hidden Markov Model

HMMs are probabilistic; they define:

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

(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.

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

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:

$$\operatorname*{argmax}_{\lambda} \mathbf{w}^\top \phi_{local}(x, L, y[L-1], \lambda) + \mathbf{w}^\top \left(\sum_{\ell=2}^{L-1} \phi_{local}(x, \ell, y[\ell-1], y[\ell]) \right)$$

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:

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

Visualization

N							
O							
$\wedge$							
V							
A							
!							
$\vdots$							
	ikr	smh	he	asked	fir	yo	...

Finding the Best Sequence y: Algorithm

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

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

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

  Store the "argmax" λ' as $B[\ell,\lambda]$.
- $y[L] = \text{argmax}_\lambda \, V[L,\lambda]$.
- *Backtrack.* For $\ell \in \{L-1,\ldots,1\}$:

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

- Return $\langle y[1],\ldots,y[L]\rangle$.

Visualizing and Analyzing Viterbi

	ikr	smh	he	asked	fir	yo	...
N							
O							
$\wedge$							
V							
A							
!							
$\vdots$							

Sequence Labeling: What's Next?

1. What is sequence labeling useful for?
2. What are the features ϕ?
3. How we learn the parameters $\mathbf{w}$?

Part-of-Speech Tagging

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	

Supersense Tagging

ikr	smh	he	asked	fir	yo	last	name
–	–	–	communication	–	–	–	cognition

so	he	can	add	u	on	fb	lololol
–	–	–	stative	–	–	group	–

See: "Coarse lexical semantic annotation with supersenses: an Arabic case study," Schneider et al. (2012).

Named Entity Recognition

With Commander Chris Ferguson at the helm ,
person

Atlantis touched down at Kennedy Space Center .
spacecraft location

Named Entity Recognition

With	Commander	Chris	Ferguson	at	the	helm	,
		person					
O	B	I	I	O	O	O	O

Atlantis	touched	down	at	Kennedy	Space	Center	.
spacecraft					location		
B	O	O	O	B	I	I	O

Named Entity Recognition: Another Example

	1	2	3	4	5	6	7	8	9	10
$\mathbf{x}$ =	Britain	sent	warships	across	the	English	Channel	Monday	to	rescue
$\mathbf{y}$ =	B	O	O	O	O	B	I	B	O	O
$\mathbf{y}'$ =	O	O	O	O	O	B	I	B	O	O

11	12	13	14	15	16	17	18	19
Britons	stranded	by	Eyjafjallajökull	's	volcanic	ash	cloud	.
B	O	O	B	O	O	O	O	O
B	O	O	B	O	O	O	O	O

Named Entity Recognition: Features

ϕ	$\phi(\mathbf{x}, \mathbf{y})$	$\phi(\mathbf{x}, \mathbf{y}')$
bias:		
count of i s.t. $y[i] = \mathrm{B}$	5	4
count of i s.t. $y[i] = \mathrm{I}$	1	1
count of i s.t. $y[i] = \mathrm{O}$	14	15
lexical:		
count of i s.t. $x[i] =$ *Britain* and $y[i] = \mathrm{B}$	1	0
count of i s.t. $x[i] =$ *Britain* and $y[i] = \mathrm{I}$	0	0
count of i s.t. $x[i] =$ *Britain* and $y[i] = \mathrm{O}$	0	1
downcased:		
count of i s.t. $lc(x[i]) =$ *britain* and $y[i] = \mathrm{B}$	1	0
count of i s.t. $lc(x[i]) =$ *britain* and $y[i] = \mathrm{I}$	0	0
count of i s.t. $lc(x[i]) =$ *britain* and $y[i] = \mathrm{O}$	0	1
count of i s.t. $lc(x[i]) =$ *sent* and $y[i] = \mathrm{O}$	1	1
count of i s.t. $lc(x[i]) =$ *warships* and $y[i] = \mathrm{O}$	1	1

Named Entity Recognition: Features

ϕ	$\phi(\mathbf{x}, \mathbf{y})$	$\phi(\mathbf{x}, \mathbf{y}')$
shape:		
count of i s.t. $shape(x[i]) = Aaaaaaa$ and $y[i] = \mathrm{B}$	3	2
count of i s.t. $shape(x[i]) = Aaaaaaa$ and $y[i] = \mathrm{I}$	1	1
count of i s.t. $shape(x[i]) = Aaaaaaa$ and $y[i] = \mathrm{O}$	0	1
prefix:		
count of i s.t. $pre_1(x[i]) = B$ and $y[i] = \mathrm{B}$	2	1
count of i s.t. $pre_1(x[i]) = B$ and $y[i] = \mathrm{I}$	0	0
count of i s.t. $pre_1(x[i]) = B$ and $y[i] = \mathrm{O}$	0	1
count of i s.t. $pre_1(x[i]) = s$ and $y[i] = \mathrm{O}$	2	2
count of i s.t. $shape(pre_1(x[i])) = A$ and $y[i] = \mathrm{B}$	5	4
count of i s.t. $shape(pre_1(x[i])) = A$ and $y[i] = \mathrm{I}$	1	1
count of i s.t. $shape(pre_1(x[i])) = A$ and $y[i] = \mathrm{O}$	0	1
$\mathbb{I}\{shape(pre_1(x[1])) = A \wedge y_1 = \mathrm{B}\}$	1	0
$\mathbb{I}\{shape(pre_1(x[1])) = A \wedge y[1] = \mathrm{O}\}$	0	1
gazetteer:		
count of i s.t. $x[i]$ is in the gazetteer and $y[i] = \mathrm{B}$	2	1
count of i s.t. $x[i]$ is in the gazetteer and $y[i] = \mathrm{I}$	0	0
count of i s.t. $x[i]$ is in the gazetteer and $y[i] = \mathrm{O}$	0	1
count of i s.t. $x[i] = sent$ and $y[i] = \mathrm{O}$	1	1

Multiword Expressions

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 I

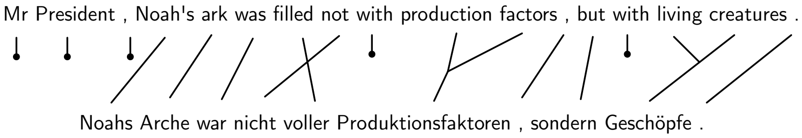
the price which means a lot to me .

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).

Cross-Lingual Word Alignment



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

	model 4	fast_align	speedup
Chinese → English	34.1	**34.7**	13×
French → English	27.4	**27.7**	10×
Arabic → English	54.5	**55.7**	10×

Other Sequence Decoding Problems

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

Add dimensions:

- Image segmentation
- Object recognition
- Optical character recognition

Sequence Decoding: L

Recall that for categorization, we set up learning as **empirical risk minimization**:

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

Example loss:

$$loss(x, y; \mathbf{w}) = -\mathbf{w}^\top \phi(x, y) + \max_{y' \in \mathcal{Y}_x} \mathbf{w}^\top \phi(x, y')$$

Structured Perceptron (Collins, 2002)

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

- $\mathbf{w} = \mathbf{0}$
- For $t \in \{1, \ldots, T\}$:
  - Draw n uniformly at random from $\{1, \ldots, N\}$.
  - Decode x_n:

$$\hat{y} = \operatorname*{argmax}_{y \in \mathcal{Y}_{x_n}} \mathbf{w}^\top \phi(x_n, y)$$

  - If $\hat{y} \neq y_n$, update parameters:

$$\mathbf{w} = \mathbf{w} + \alpha_t \left(\phi(x_n, y_n) - \phi(x_n, \hat{y}) \right)$$

- Return $\mathbf{w}$

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 $\Omega(\mathbf{w})$, as previously discussed.

Change the optimization algorithm:

- Automatic step-size scaling (e.g., MIRA, Adagrad)
- Batch and "mini-batch" updating
- Averaging and voting

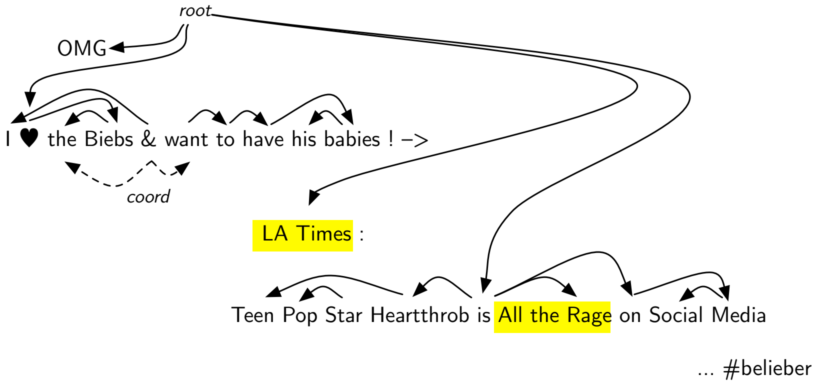
Structured Prediction: Lines of Attack

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

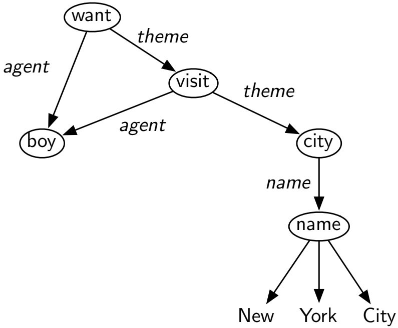
- Can all linguistic structure be captured with sequence labeling?
- Some representations are more elegantly handled using other kinds of output structures.
  - Syntax: trees
  - Semantics: graphs
- Dynamic programming and other combinatorial algorithms are central.
  - Always useful: features ϕ that decompose into local parts

Dependency Tree



See: "A dependency parser for tweets," Kong et al. (2014)

Semantic Graph



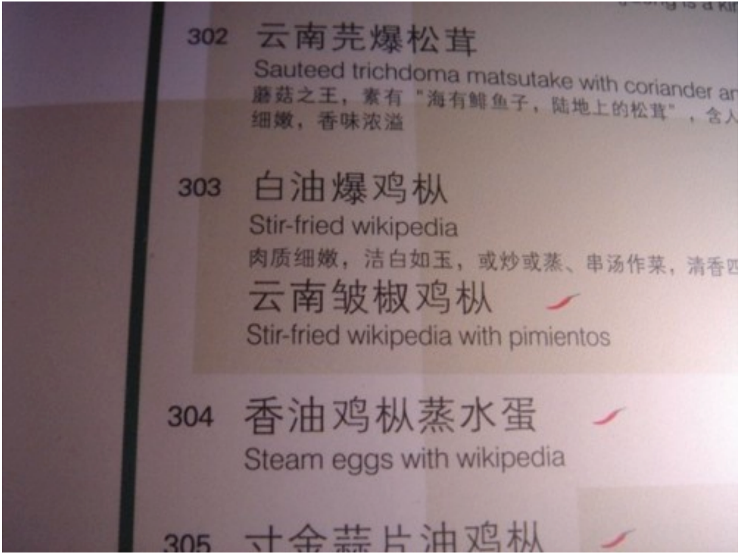
The boy wants to visit New York City.

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

Part III

Example Applications

Machine Translation



Translation from Analytic to Synthetic Languages

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

Useful tool: morphological lexicon

$y_\sigma =$ пытаться

$y_\mu = \{$Verb, MAIN, IND, PAST, SING, FEM, MEDIAL, PERF$\}$

deterministic

пыталась

"Translating into morphologically rich languages with synthetic phrases," Chahuneau et al. (2013)

High-Level Approach

Contemporary translation is performed by mapping source-language "phrases" to target-language "phrases."

A phrase is a sequence of one or more words.

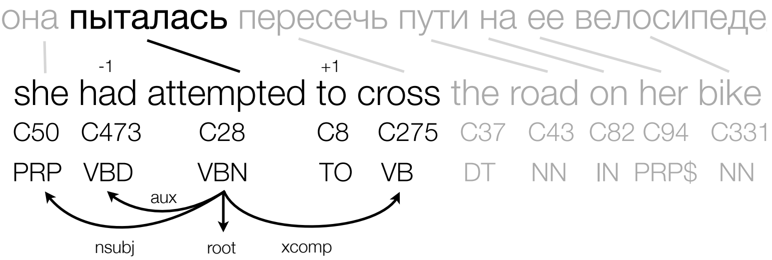
In addition, let a phrase be a sequence of one or more *stems*.

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

Predicting Inflection in Multilingual Context

y_σ = пытаться

$y_\mu = \{$Verb, MAIN, IND, PAST, SING, FEM, MEDIAL, PERF$\}$



$$\phi(x, y_\mu) = \langle \phi_{source}(x) \otimes \phi_{target}(y_\mu), \phi_{target}(y_\mu) \otimes \phi_{target}(y_\mu) \rangle$$

Translation Results (out of English)

	→ Russian	→ Hebrew	→ Swahili
Baseline	14.7$\pm$0.1	15.8$\pm$0.3	18.3$\pm$0.1
+Class LM	15.7$\pm$0.1	16.8$\pm$0.4	18.7$\pm$0.2
+Synthetic	16.2$\pm$0.1	17.6$\pm$0.1	19.0$\pm$0.1

Translation quality (Bleu score; higher is better), averaged across three runs.

Something Completely Different

Measuring Ideological Proportions

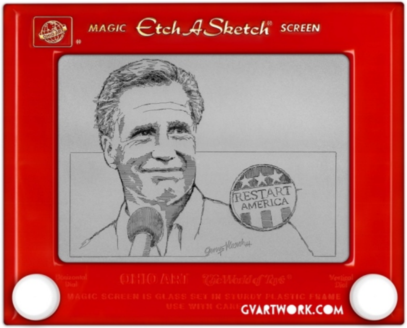
"Well, I think you hit a reset button for the fall campaign. Everything changes. It's almost like an Etch-A-Sketch. You can kind of shake it up and restart all over again."

—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



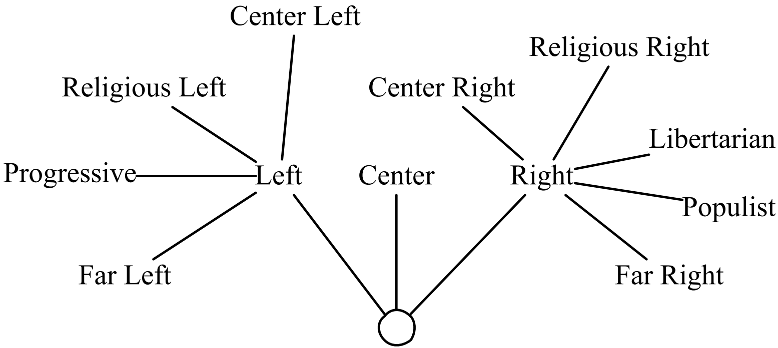
Measuring Ideological Proportions: Motivation

- 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”



Cue-Lag Representation of a Speech

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

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

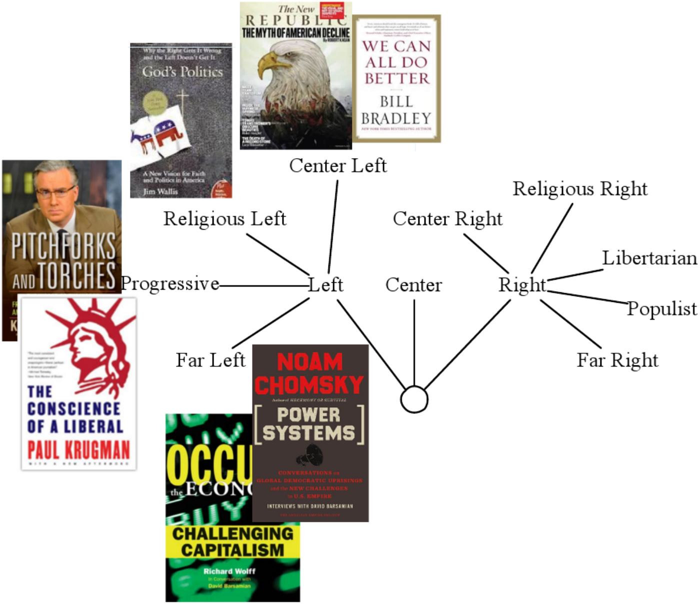
Social Security 24 tax increase 13 growth rate 21 higher taxes 29

health care costs

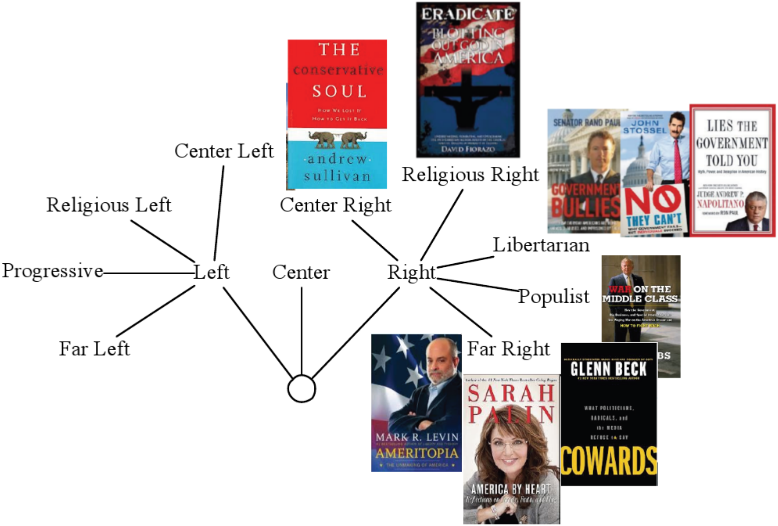
Line of Attack

1. Build a "dictionary" of cues.
2. Infer ideological proportions from the cue-lag representation of speeches.

Ideological Books Corpus



Ideological Books Corpus



Example Cues

Center-Right D. Frum, M. McCain, C. T. Whitman (1,450)	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
Libertarian Rand Paul, John Stossel, *Reason* (2,268)	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
Religious Right (960)	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

Browse results at `http://www.ark.cs.cmu.edu/CLIP/`.

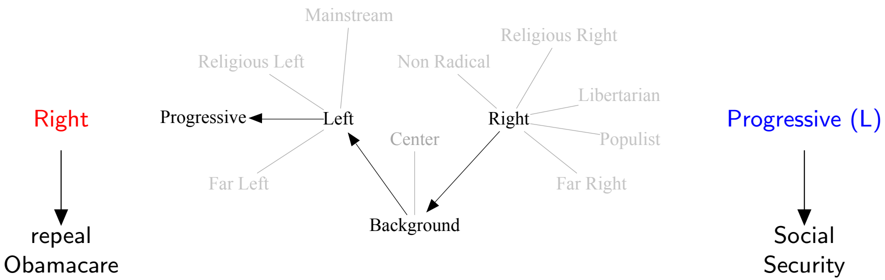
Cue-Lag Ideological Proportions Model

$$
\begin{tikzcd}
\text{Libertarian (R)} \arrow[r] \arrow[d] & \text{Libertarian (R)} \arrow[r] \arrow[d] & \text{Right} \arrow[r] \arrow[d] & \text{Progressive (L)} \arrow[d] \\
\text{government spending} & \text{federal spending} & \text{repeal Obamacare} & \text{Social Security}
\end{tikzcd}
$$

- Each speech is modeled as a sequence:
  - ideologies are labels (y)
  - cue terms are observed (x)

HMM “with a Twist”

HMM "with a Twist"



$$\mathbf{w}^\top \phi_{local}(x, \ell, \text{Right}, \text{Prog.}) = \log p(\text{Right} \rightsquigarrow \text{Prog.}) + \ldots$$

HMM "with a Twist"

Right → repeal Obamacare

lag=7

Progressive (L) → Social Security

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."

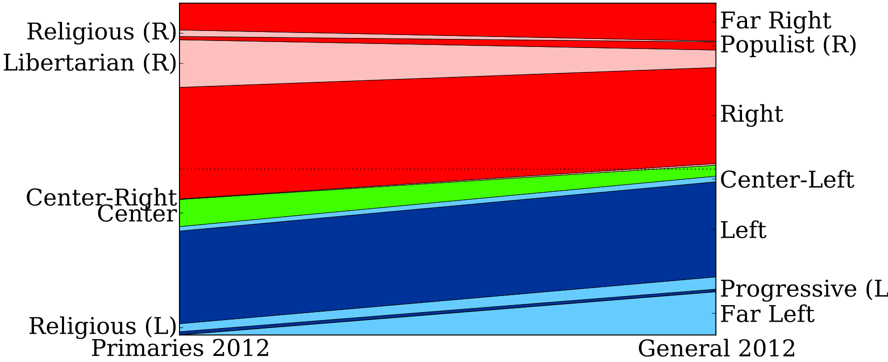
We sample from the posterior over labels, $p(y \mid x)$.

This is sometimes called *approximate Bayesian inference*.

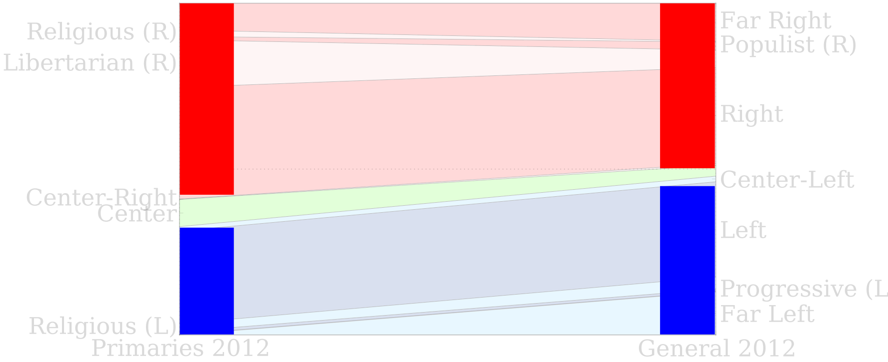
Measuring Ideological Proportions in Speeches

- 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

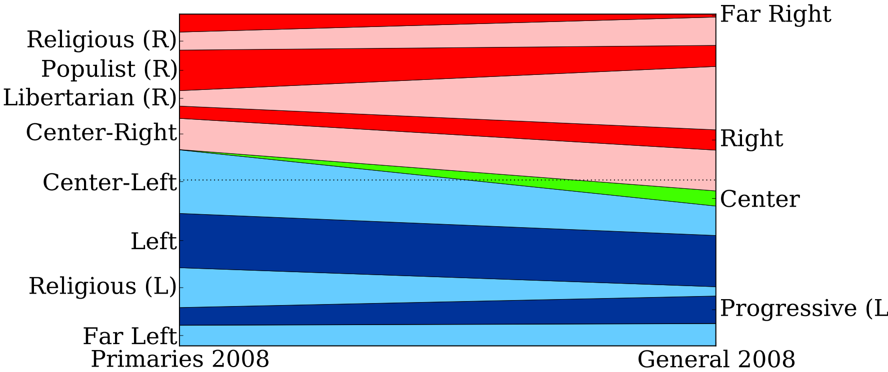

Religious (R)
Libertarian (R)
Center-Right
Center
Religious (L)
Far Right
Populist (R)
Right
Center-Left
Left
Progressive (L
Far Left
Primaries 2012
General 2012

Mitt Romney

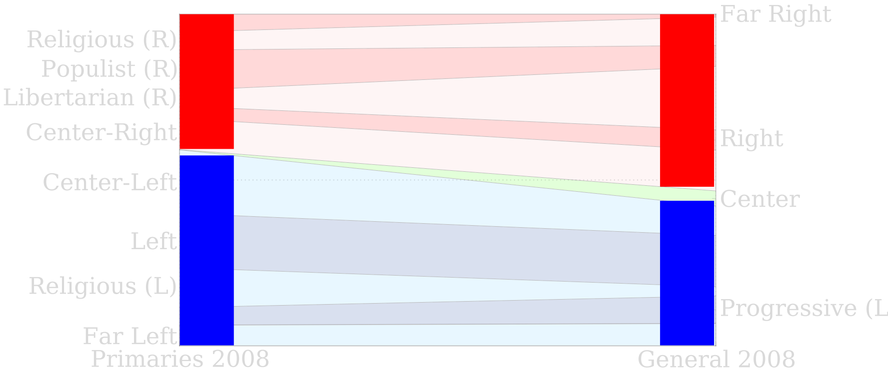


Religious (R)
Libertarian (R)
Center-Right
Center
Religious (L)
Primaries 2012
Far Right
Populist (R)
Right
Center-Left
Left
Progressive (L
Far Left
General 2012

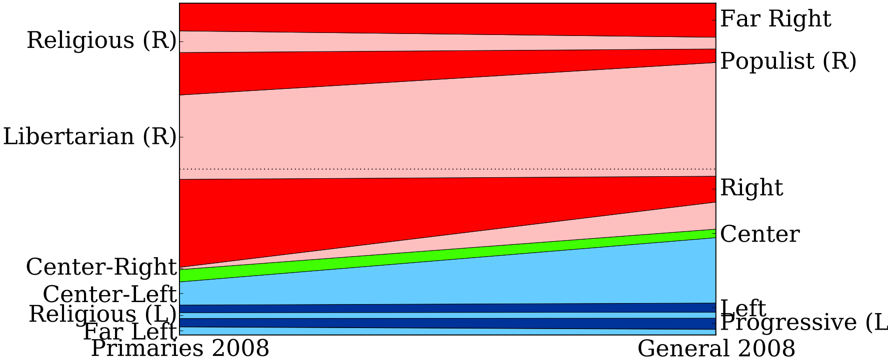
Barack Obama



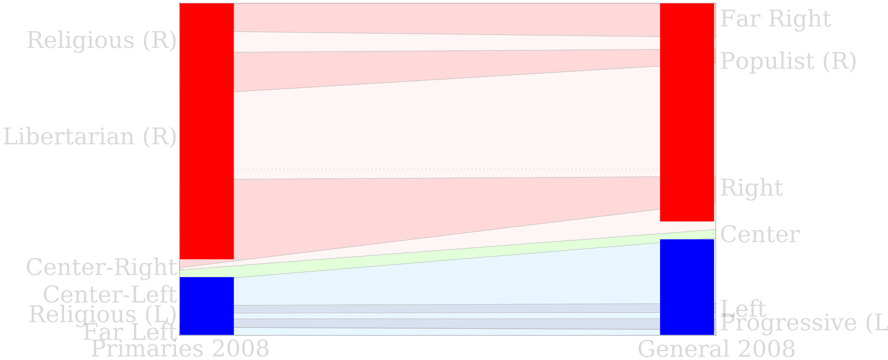
Barack Obama



John McCain


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

John McCain


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

Objective Evaluation?

Pre-registered hypothesis

A statement by a domain expert about his/her *expectations* of the model's output.

Preregistered Hypotheses

Hypotheses	
Sanity checks (strong):	
S1.	Republican primary candidates should tend to draw more from RIGHT than from LEFT.
S2.	Democratic primary candidates should tend to draw more from LEFT than from RIGHT.
S3.	In general elections, Democrats should draw more from the LEFT than the Republicans and vice versa for the RIGHT.
Primary hypotheses (strong):	
P1.	Romney, McCain and other Republicans should almost never draw from FAR LEFT, and extremely rarely from PROGRESSIVE.
P2.	Romney should draw more heavily from the RIGHT than Obama in both stages of the 2012 campaign.
Primary hypotheses (moderate):	
P3.	Romney should draw more heavily on words from the LIBERTARIAN, POPULIST, RELIGIOUS RIGHT, and FAR RIGHT in the primary compared to the general election. In the general election, Romney should draw more heavily on CENTER, CENTER-RIGHT and LEFT vocabularies.

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

	CLIP	HMM	Mix	NoRes
Sanity checks	**20/21**	19/22	21/22	17/22
Strong hypotheses	**31/34**	23/33	28/34	30/34
Moderate hypotheses	**14/17**	**14/17**	12/17	11/17
Total	**65/72**	56/72	61/73	58/73

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 I

Blitzer, J., Dredze, M., and Pereira, F. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Chahuneau, V., Schlinger, E., Dyer, C., and Smith, N. A. (2013). Translating into morphologically rich languages with synthetic phrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Collins, M. (2002). Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Flanigan, J., Thomson, S., Carbonell, J., Dyer, C., and Smith, N. A. (2014). A discriminative graph-based parser for the abstract meaning representation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N. A. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, companion volume*.

Kong, L., Schneider, N., Swayamdipta, S., Bhatia, A., Dyer, C., and Smith, N. A. (2014). A dependency parser for tweets. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Martins, A. F. T., Yogatama, D., Smith, N. A., and Figueiredo, M. A. T. (2014). Structured sparsity in natural language processing: Models, algorithms, and applications. EACL tutorial available at `http://www.cs.cmu.edu/~afm/Home_files/eacl2014tutorial.pdf`.

Mosteller, F. and Wallace, D. L. (1963). Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed Federalist Papers. *Journal of the American Statistical Association*, 58(302):275–309.

Owoputi, O., O'Connor, B., Dyer, C., Gimpel, K., Schneider, N., and Smith, N. A. (2013). Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

References II

Schneider, N., Danchik, E., Dyer, C., and Smith, N. A. (2014). Discriminative lexical semantic segmentation with gaps: Running the MWE gamut. *Transactions of the Association for Computational Linguistics*, 2:193–206.

Schneider, N., Mohit, B., Oflazer, K., and Smith, N. A. (2012). Coarse lexical semantic annotation with supersenses: An Arabic case study. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Sim, Y., Acree, B. D. L., Gross, J. H., and Smith, N. A. (2013). Measuring ideological proportions in political speeches. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Seattle, WA.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288.

Yano, T., Smith, N. A., and Wilkerson, J. D. (2012). Textual predictors of bill survival in Congressional committees. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*.

Yogatama, D. and Smith, N. A. (2014). Making the most of bag of words: Sentence regularization with alternating direction method of multipliers. In *Proceedings of the International Conference on Machine Learning*.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society (B)*, 68(1):49.