Rich Prior Knowledge in Learning for NLP

Gregory Druck, Kuzman Ganchev, João Graça

updated slides: http://sideinfo.wikii.com
Goal: Build a Statistical NLP System

updated slides: http://sideinfo.wikii.com
Goal: Build a Statistical NLP System

updated slides: http://sideinfo.wikii.com
Goal: Build a Statistical NLP System

have: unlabeled data

What are our options?

updated slides: http://sideinfo.wikii.com
Option 1: Label Data

have: unlabeled data

updated slides: http://sideinfo.wikii.com
Option 1: Label Data

have: unlabeled data

hire: linguist
Option 1: Label Data

**have:** unlabeled data

**hire:**
- linguist
- annotators

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Option 1: Label Data

**have:** unlabeled data

**hire:** $$$

linguist annotators

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Option 1: Label Data

have: unlabeled data
Option 1: Label Data

have: unlabeled data

hire: $$$$$$

linguist  annotators

updated slides: http://sideinfo.wikkii.com
Option 1: Label Data

have: unlabeled data

updated slides: http://sideinfo.wikii.com
Option 1: Label Data

have: unlabeled data

hire: $$$$$$$$$$$$$$$$

linguist

annotators

updated slides: http://sideinfo.wikii.com
Option 1: Label Data

This approach does not scale to every task and domain of interest.

have: unlabeled data

hire: $$$$$$$$$$$$$$$$

linguist

annotators

updated slides: http://sideinfo.wikii.com
Option II: Unsupervised Learning

*have:* unlabeled data

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Option II: Unsupervised Learning

**Have:** unlabeled data

**Design:** model

\[ y_1 \rightarrow y_2 \rightarrow y_3 \]

\[ x_1 \rightarrow x_2 \rightarrow x_3 \]
Option II: Unsupervised Learning

**have:** unlabeled data

**design:** model

**train:** to maximize likelihood of observed data
Option II: Unsupervised Learning

The true generative process is typically:

- unknown
- complex; hard to model efficiently

**Have:** unlabeled data

**Design:** model

**Train:** to maximize likelihood of observed data

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Option II: Unsupervised Learning

the true generative process is typically:

- unknown
- complex; hard to model efficiently

result: maximizing likelihood may not give expected output ...

have: unlabeled data

design: model

train: to maximize likelihood of observed data

jugaban de una manera animada y muy cordial

it was an animated, very convivial game

updated slides: http://sideinfo.wikii.com
Prior Knowledge

We posses a wealth of prior knowledge about most NLP tasks.
Example: Document Classification

Documents

Labels
Example: Document Classification

- **Prior Knowledge:**
  - labeled features: information about the labels for documents that contain a particular word $w$
Example: Document Classification

- **Prior Knowledge:**

- labeled features: information about the labels for documents that contain a particular word w

<table>
<thead>
<tr>
<th>sentiment polarity</th>
<th>newsgroups classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>positive</td>
<td>baseball</td>
</tr>
<tr>
<td>memorable</td>
<td>hit</td>
</tr>
<tr>
<td>perfect</td>
<td>Braves</td>
</tr>
<tr>
<td>exciting</td>
<td>runs</td>
</tr>
<tr>
<td>negative</td>
<td>Mac</td>
</tr>
<tr>
<td>terrible</td>
<td>Apple</td>
</tr>
<tr>
<td>boring</td>
<td>Macintosh</td>
</tr>
<tr>
<td>mess</td>
<td>Powerbook</td>
</tr>
<tr>
<td></td>
<td>politics</td>
</tr>
<tr>
<td></td>
<td>senate</td>
</tr>
<tr>
<td></td>
<td>taxes</td>
</tr>
<tr>
<td></td>
<td>liberal</td>
</tr>
</tbody>
</table>

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Example: Information Extraction


updated slides: http://sideinfo.wikkii.com
Example: Information Extraction

Example: Information Extraction

- Prior Knowledge:


updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Example: Information Extraction

- Prior Knowledge:
  - labeled features:
    - the word **ACM** should be labeled either *journal* or *conference* most of the time

---


---

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Example: Information Extraction

Extraction from research papers:


Prior Knowledge:

- labeled features:
  - the word **ACM** should be labeled either *journal* or *conference* most of the time

- non-Markovian (long-range) dependencies:
  - each reference has at most one segment of each type

Updated slides: http://sideinfo.wikkii.com
Example: Part-of-speech Induction

A career with the European institutions must become more attractive. Too many young, new...
Example: Part-of-speech Induction

A career with the European institutions must become more attractive. Too many young, new...
Example: Part-of-speech Induction

A career with the European institutions must become more attractive. Too many young, new...

• Prior Knowledge:
Example: Part-of-speech Induction

Prior Knowledge:

• linguistic knowledge: each sentence should have a verb

Tags

Text

A career with the European institutions must become more attractive. Too many young, new...
Example: Part-of-speech Induction

Tags

Text

A career with the European institutions must become more attractive. Too many young, new...

• **Prior Knowledge:**
  
  • *linguistic knowledge*: *each sentence should have a verb*
  
  • *linguistic knowledge*: *the total number of different POS tags assigned to each word type should be small*

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Example: Dependency Grammar Induction

root → John → hit → the → ball → with → the → bat
Example: Dependency Grammar Induction

root → John → hit → the → ball → with → the → bat
Example: Dependency Grammar Induction

- **Prior Knowledge:**

```plaintext
root  John  hit  the  ball  with  the  bat
```
Example: Dependency Grammar Induction

- **Prior Knowledge:**
  - linguistic rules: *nouns are usually dependents of verbs*
Example: Dependency Grammar Induction

- **Prior Knowledge:**
  - *linguistic rules:* nouns are usually dependents of verbs
  - *parallel corpora:* target language parses should be similar to aligned parses in a resource-rich source language
Example: Word Alignment

A career with the European institutions must become more attractive.

Uma carreira nas instituições europeias têm de se tornar mais atractiva.
Example: Word Alignment

A career with the European institutions must become more attractive.

Uma carreira nas instituições europeias têm de se tornar mais atractiva.
Example: Word Alignment

A career with the European institutions must become more attractive.

Uma carreira nas instituições europeias tem de se tornar mais atrativa.

- Prior Knowledge:
Example: Word Alignment

A career with the European institutions must become more attractive.

Uma carreira nas instituições europeias tem de se tornar mais atractiva.

- **Prior Knowledge:**
- **Bijectivity**: *alignment should be mostly one-to-one*
Example: Word Alignment

A career with the European institutions must become more attractive.

Uma carreira nas instituições europeias tem de se tornar mais atractiva.

- **Prior Knowledge:**
  - **Bijectivity:** *alignment should be mostly one-to-one*
  - **Symmetry:** *source* $\rightarrow$ *target* and *target* $\rightarrow$ *source* alignments should agree
This Tutorial

In general, how can we leverage such knowledge and an unannotated corpus during learning?
Tutorial Organization
Tutorial Organization

- Motivation & Introduction [Greg]
Tutorial Organization

• **Motivation & Introduction** [Greg]

• **Frameworks & Connections** [Kuzman]

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Tutorial Organization

- **Motivation & Introduction** [Greg]
- **Frameworks & Connections** [Kuzman]
- **Survey of Applications** [João]

updated slides: http://sideinfo.wikkii.com
Tutorial Organization

- **Motivation & Introduction** [Greg]
- **Frameworks & Connections** [Kuzman]
- **Survey of Applications** [João]
- **Implementation** [Greg]

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
### Notation & Models

<table>
<thead>
<tr>
<th>Input variables (documents, sentences):</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structured output variables (parses, sequences):</td>
<td>$y$</td>
</tr>
<tr>
<td>Unstructured output variables (labels):</td>
<td>$y$</td>
</tr>
<tr>
<td>Input / output variables for entire corpus:</td>
<td>$X$ $Y$</td>
</tr>
<tr>
<td>Probabilistic model parameters:</td>
<td>$\theta$</td>
</tr>
<tr>
<td>Generative models:</td>
<td>$p_\theta(x, y)$</td>
</tr>
<tr>
<td>Discriminative models:</td>
<td>$p_\theta(y</td>
</tr>
<tr>
<td>Model feature function:</td>
<td>$f(x, y)$</td>
</tr>
</tbody>
</table>
Running Example #1:
Document Classification

updated slides: http://sideinfo.wikii.com
Running Example #1: Document Classification

- **model**: Maximum Entropy Classifier (Logistic Regression)
Running Example #1:
Document Classification

- **model:** Maximum Entropy Classifier (Logistic Regression)

\[ p_\theta(y|x) = \frac{1}{Z(x)} \exp(\theta \cdot f(x, y)) \]
Running Example #1:
Document Classification

- **model:** Maximum Entropy Classifier (Logistic Regression)

\[
p_\theta(y|x) = \frac{1}{Z(x)} \exp(\theta \cdot f(x, y))
\]
Running Example #1: Document Classification

- **model**: Maximum Entropy Classifier (Logistic Regression)

\[ p_{\theta}(y|x) = \frac{1}{Z(x)} \exp(\theta \cdot f(x, y)) \]

updated slides: http://sideinfo.wikii.com
Running Example #1: Document Classification

- **model:** Maximum Entropy Classifier (Logistic Regression)

\[ p_\theta(y|x) = \frac{1}{Z(x)} \exp(\theta \cdot f(x, y)) \]
Running Example #1: Document Classification

- **model**: Maximum Entropy Classifier (Logistic Regression)

\[
p_\theta(y|x) = \frac{1}{Z(x)} \exp(\theta \cdot f(x, y))
\]

- **no labeled documents**
Running Example #1: Document Classification

- **model**: Maximum Entropy Classifier (Logistic Regression)

\[
p_\theta(y|x) = \frac{1}{Z(x)} \exp(\theta \cdot f(x, y))
\]

- *no labeled documents*

- **prior knowledge**:
Running Example #1:
Document Classification

• **model**: Maximum Entropy Classifier (Logistic Regression)

\[
p_\theta(y|x) = \frac{1}{Z(x)} \exp(\theta \cdot f(x, y))
\]

• **no labeled documents**

• **prior knowledge**: 
  • labeled features: information about the label distribution when word \(w\) is present
Running Example #1: Document Classification

- **model:** Maximum Entropy Classifier (Logistic Regression)

\[
p_{\theta}(y|x) = \frac{1}{Z(x)} \exp(\theta \cdot f(x, y))
\]

- **no labeled documents**

- **prior knowledge:**
  - labeled features: information about the label distribution when word \( w \) is present
  - label is often **hockey** or **baseball** when **game** is present
Running Example #2: Word Alignment

updated slides: http://sideinfo.wikkii.com
Running Example #2: Word Alignment

- **model**: first-order Hidden Markov Model (HMM)
Running Example #2: Word Alignment

- **model**: first-order Hidden Markov Model (HMM)

\[
p_\theta(y, x) = p_\theta(y_0) \prod_{i=1}^{N} p_\theta(y_i|y_{i-1}) p_\theta(x_i|y_i)
\]

![Diagram of word alignment](image)
Running Example #2: Word Alignment

- **model:** first-order Hidden Markov Model (HMM)

\[ p_\theta(y, x) = p_\theta(y_0) \prod_{i=1}^{N} p_\theta(y_i|y_{i-1}) p_\theta(x_i|y_i) \]
Running Example #2: Word Alignment

• **model**: first-order Hidden Markov Model (HMM)

\[
p_{\theta}(y, x) = p_{\theta}(y_0) \prod_{i=1}^{N} p_{\theta}(y_i | y_{i-1}) p_{\theta}(x_i | y_i)
\]

output (alignment)  input (sentences)

\[
\begin{align*}
1 & \quad \text{sabemos} \\
2 & \quad \text{know} \\
3 & \quad \text{the} \\
0 & \quad \text{null}
\end{align*}
\]

\[
\begin{align*}
1 & \quad \text{we} \\
2 & \quad \text{el} \\
3 & \quad \text{camino} \\
0 & \quad \text{null}
\end{align*}
\]

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Running Example #2:
Word Alignment

• **model:** first-order Hidden Markov Model (HMM)

\[
p_{\theta}(y, x) = p_{\theta}(y_0) \prod_{i=1}^{N} p_{\theta}(y_i|y_{i-1})p_{\theta}(x_i|y_i)
\]

• *no annotated alignments*
Running Example #2: Word Alignment

- **model**: first-order Hidden Markov Model (HMM)

\[
p_\theta(y, x) = p_\theta(y_0) \prod_{i=1}^{N} p_\theta(y_i|y_{i-1})p_\theta(x_i|y_i)
\]

- **no annotated alignments**
- **prior knowledge:**

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Running Example #2: Word Alignment

- **model:** first-order Hidden Markov Model (HMM)

\[ p_\theta(y, x) = p_\theta(y_0) \prod_{i=1}^{N} p_\theta(y_i | y_{i-1}) p_\theta(x_i | y_i) \]

- **no annotated alignments**

- **prior knowledge:**
  - **Bijectivity:** alignment should be mostly one-to-one
Example #2: Without Prior Knowledge
Example #2: Without Prior Knowledge

HMM

updated slides: http://sideinfo.wikii.com
Example #2: Without Prior Knowledge

HMM

sentences

updated slides: http://sideinfo.wikii.com
Example #2: Without Prior Knowledge

HMM

\[
\begin{align*}
&y_1 \quad y_2 \quad y_3 \\
&x_1 \quad x_2 \quad x_3 
\end{align*}
\]

sentences

\[ + \]

output

jugaban de una manera animada y muy cordial

it was an animated, very convivial game

updated slides: http://sideinfo.wikii.com
Example #2: Without Prior Knowledge

HMM  sentences  output

jugaban de una manera animada y muy cordial
it was an animated, very convivial game

This output does not agree with prior knowledge!

- six target words align to source word *animada*
- five source words do not align with any target word

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Leveraging Prior Knowledge

Possible approaches and their limitations.
Limited Approach: Labeling Data

**approach:** Use prior knowledge to label data.
Limited Approach: Labeling Data

**approach**: Use *prior knowledge* to label data.

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Limited Approach: Labeling Data

**approach:** Use *prior knowledge* to label data.

Prototypes (+ cluster features):
- [Haghighi & Klein 06]

Others:
- [Raghavan & Allan 07]
- [Schapire et al. 02]

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Limited Approach: Labeling Data

**approach:** Use prior knowledge to label data.

**limitation:** Often unclear how to label data.

Prototypes (+ cluster features):
- [Haghighi & Klein 06]

Others:
- [Raghavan & Allan 07]
- [Schapire et al. 02]

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
Limited Approach: Labeling Data

**Approach:** Use prior knowledge to label data.

- Prototypes (+ cluster features):
  - [Haghighi & Klein 06]
- Others:
  - [Raghavan & Allan 07]
  - [Schapire et al. 02]

**Limitation:** Often unclear how to label data.

- **Example #1:** often (not always) *game* $\rightarrow$ \{hockey, baseball\}

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Limited Approach: Labeling Data

**approach:** Use *prior knowledge* to label data.

**limitation:** Often unclear how to label data.

- **Example #1:** often (not always) *game* $\rightarrow \{\text{hockey, baseball}\}

- **Example #2:** *alignment* should be mostly one-to-one

Prototypes (+ cluster features):
- [Haghighi & Klein 06]

Others:
- [Raghavan & Allan 07]
- [Schapire et al. 02]

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
Limited Approach: Bayesian Approach

**approach:** Encode prior knowledge with a prior on parameters.
Limited Approach: Bayesian Approach

**approach:** Encode prior knowledge with a prior on parameters.

**specifying** $p(\theta)$

**natural:** “$\theta$ should be small (or sparse)”

[Johnson 07], among many others

**possible:** “$\theta_i$ should be close to $\tilde{\theta}_i$.”

( informative prior ) [Dayanik et al. 06]
Limited Approach: Bayesian Approach

**approach:** Encode prior knowledge with a prior on parameters.

**specifying** $p(\theta)$

**natural:** “$\theta$ should be small (or sparse)”

[Johnson 07], among many others

**possible:** “$\theta_i$ should be close to $\tilde{\theta}_i$”

( informative prior ) [Dayanik et al. 06]

**limitation:** Our prior knowledge is not about parameters!
Parameters are difficult to interpret; hard to get desired effect.

updated slides: http://sideinfo.wikii.com
Limited Approach: Bayesian Approach

**approach:** Encode prior knowledge with a prior on parameters.

- specifying $p(\theta)$
- natural: “$\theta$ should be small (or sparse)”
- possible: “$\theta_i$ should be close to $\tilde{\theta}_i$”
  
  [Johnson 07], among many others

**limitation:** Our prior knowledge is not about parameters!
Parameters are difficult to interpret; hard to get desired effect.

- **Example #1:** often (not always) game $\rightarrow$ \{hockey, baseball\}

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Limited Approach: Bayesian Approach

**approach:** Encode prior knowledge with a prior on parameters.

specifying $p(\theta)$

natural: “$\theta$ should be small (or sparse)”

[Johsnon 07], among many others

possible: “$\theta_i$ should be close to $\tilde{\theta}_i$”

(informative prior) [Dayanik et al. 06]

**limitation:** Our prior knowledge is not about parameters! Parameters are difficult to interpret; hard to get desired effect.

- **Example #1:** often (not always) game $\rightarrow$ {hockey, baseball}
- **Example #2:** alignment should be mostly one-to-one

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com) 23
Limited Approach: Augmenting Model

**approach:** Encode prior knowledge with additional variables and dependencies.

[Li 2009], (arguably) many unsupervised methods
Limited Approach: Augmenting Model

**approach:** Encode prior knowledge with additional variables and dependencies.

[Li 2009], (arguably) many unsupervised methods

**limitation:** can be difficult to get desired effect

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Limited Approach: Augmenting Model

**approach:** Encode prior knowledge with additional variables and dependencies.

[Li 2009], (arguably) many unsupervised methods

**limitation:** can be difficult to get desired effect

- **Example #1:** often (not always) game → \{hockey, baseball\}
Limited Approach: Augmenting Model

**approach:** Encode prior knowledge with additional variables and dependencies.

[Li 2009], (arguably) many unsupervised methods

**limitation:** can be difficult to get desired effect

- **Example #1:** often (not always) game $\rightarrow \{\text{hockey}, \text{baseball}\}$

**limitation:** may make exact inference intractable

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Limited Approach: Augmenting Model

**approach:** Encode prior knowledge with additional variables and dependencies.

[Li 2009], (arguably) many unsupervised methods

**limitation:** can be difficult to get desired effect

- **Example #1:** often (not always) *game* $\rightarrow \{\text{hockey, baseball}\}$

**limitation:** may make exact inference intractable

- **Example #2:** Bijectivity makes inference $\#P$-complete

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
How can we address these limitations?
This Tutorial
This Tutorial

develop:
This Tutorial

develop:

• a language for directly expressing prior knowledge
This Tutorial

develop:

• a **language** for *directly* expressing prior knowledge

• **methods for learning** with knowledge in this language
This Tutorial

develop:

- a **language** for *directly* expressing prior knowledge
- **methods for learning** with knowledge in this language
  - (approximations to modeling this language directly)
This Tutorial

develop:

• a language for directly expressing prior knowledge

• methods for learning with knowledge in this language
  • (approximations to modeling this language directly)

• (loosely) these methods perform mappings for us:
This Tutorial

develop:
• a language for directly expressing prior knowledge
• methods for learning with knowledge in this language
  • (approximations to modeling this language directly)
• (loosely) these methods perform mappings for us:
  • expressed prior knowledge $\rightsquigarrow$ parameters $\theta$
This Tutorial

develop:

• a **language** for *directly* expressing prior knowledge

• **methods for learning** with knowledge in this language

  • (approximations to modeling this language directly)

• (loosely) these methods **perform mappings for us**:

  • expressed prior knowledge $\overset{\sim}{\rightarrow}$ parameters $\theta$

  • expressed prior knowledge $\overset{\sim}{\rightarrow}$ labeling

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com) 26
A Language for Encoding Prior Knowledge

updated slides: http://sideinfo.wikkii.com
A Language for Encoding Prior Knowledge

Our prior knowledge is about distributions over latent output variables. (output variables are interpretable)
A Language for Encoding Prior Knowledge

Our prior knowledge is about distributions over latent output variables. (output variables are interpretable)

We know some properties of this distribution:
Our prior knowledge is about distributions over latent output variables. (output variables are interpretable)

We know some properties of this distribution:

- **Example #1:** often (not always) game → {hockey, baseball}
A Language for Encoding Prior Knowledge

Our prior knowledge is about **distributions over latent output variables**. (output variables are interpretable)

We know some *properties* of this distribution:

- **Example #1**: often (not always) \texttt{game} → \{hockey, baseball\}

\begin{itemize}
  \item \texttt{baseball} \hspace{1cm} \texttt{hockey} \hspace{1cm} \texttt{politics} \hspace{1cm} \texttt{science}
\end{itemize}
A Language for Encoding Prior Knowledge

Our prior knowledge is about distributions over latent output variables. (output variables are interpretable)

We know some properties of this distribution:

- **Example #1:** often (not always) game → {hockey, baseball}

```
baseball  hockey  politics  science
```

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
A Language for Encoding Prior Knowledge

Our prior knowledge is about distributions over latent output variables. (output variables are interpretable)

We know some properties of this distribution:

• Example #1: often (not always) game → {hockey, baseball}

baseball hockey politics science

contain game

updated slides: http://sideinfo.wikkii.com
A Language for Encoding Prior Knowledge

Our prior knowledge is about **distributions over latent output variables**. (output variables are interpretable)

We know some *properties* of this distribution:

- **Example #1:** often (not always) \( \text{game} \rightarrow \{ \text{hockey, baseball} \} \)

```plaintext
baseball  hockey  politics  science
```

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
A Language for Encoding Prior Knowledge

Our prior knowledge is about \textit{distributions over latent output variables}. (output variables are interpretable)

We know some \textit{properties} of this distribution:

- \textbf{Example \#1:} often (not always) \texttt{game} $\rightarrow$ \{hockey, baseball\}

---

baseball  hockey  politics  science

---

contain \texttt{game}

contain \texttt{game}

updated slides: \url{http://sideinfo.wikkii.com}  27
A Language for Encoding Prior Knowledge

Our prior knowledge is about **distributions over latent output variables**. (output variables are interpretable)

We know some *properties* of this distribution:

- **Example #1:** often (not always) $\text{game} \rightarrow \{\text{hockey}, \text{baseball}\}$

The slides contain:

- baseball
- hockey
- politics
- science

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
A Language for Encoding Prior Knowledge

- **Formulation:** know about the **expectations** of some functions under distribution over latent output variables
A Language for Encoding Prior Knowledge

- **Formulation**: know about the **expectations** of some functions under distribution over latent output variables

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
A Language for Encoding Prior Knowledge

- **Formulation:** know about the **expectations** of some functions under distribution over latent output variables

expected label distributions for documents with *game*

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
A Language for Encoding Prior Knowledge

- **Formulation:** know about the *expectations* of some functions under distribution over latent output variables

  expected label distributions for documents with *game*

  contain *game*  >  contain *game*

  40.0% hockey  
  43.3% baseball  
  16.7% other

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
A Language for Encoding Prior Knowledge

• **Formulation:** know about the *expectations* of some functions under distribution over latent output variables

expected label distributions for documents with *game*

<table>
<thead>
<tr>
<th>contain game</th>
<th>contain game</th>
</tr>
</thead>
<tbody>
<tr>
<td>40.0% hockey</td>
<td>16.7% hockey</td>
</tr>
<tr>
<td>43.3% baseball</td>
<td>0.0% baseball</td>
</tr>
<tr>
<td>16.7% other</td>
<td>83.3% other</td>
</tr>
</tbody>
</table>

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Constraint Features & Expectations: Document Classification

- **constraint feature:**

\[ \phi_{w \ell}(x, y) = \begin{cases} 
1 & \text{if game is in } x \text{ and } y \text{ is hockey} \\
0 & \text{otherwise} 
\end{cases} \]
Constraint Features & Expectations: Document Classification

- **constraint feature:**
  \[
  \phi_{w\ell}(x, y) = \begin{cases} 
  1 & \text{if } game \text{ is in } x \text{ and } y \text{ is hockey} \\
  0 & \text{otherwise}
  \end{cases}
  \]

- **expectation:**
  \[
  E_{p_{\theta}}[\phi_{w\ell}(X, Y)] = \frac{1}{c_w} \sum_{x} \sum_{y} p_{\theta}(y|x) \phi_{w\ell}(x, y)
  \]

- **expected** probability that documents that contain \(game\) are labeled \textit{hockey} \((c_w \text{ is the count of } game)\)
Constraint Features & Expectations: Document Classification

- **Constraint feature:**

\[ \phi_{w\ell}(x, y) = \begin{cases} 
1 & \text{if } game \text{ is in } x \text{ and } y \text{ is hockey} \\
0 & \text{otherwise} 
\end{cases} \]

- **Expectation:**

\[ \mathbb{E}_{p\theta}[\phi_{w\ell}(X, Y)] = \frac{1}{c_w} \sum_{x} \sum_{y} p(y|x) \phi_{w\ell}(x, y) \]

- **Expected** probability that documents that contain *game* are labeled *hockey* (*c_w* is the count of *game*)

**labels**
- baseball
- hockey
- politics
- science

**contain game**

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Constraint Features & Expectations: Document Classification

- **constraint feature:**
  \[ \phi_{\ell}(x, y) = \begin{cases} 1 & \text{if game is in } x \text{ and } y \text{ is hockey} \\ 0 & \text{otherwise} \end{cases} \]

- **expectation:**
  \[ E_{p_\theta}[\phi_{\ell}(X, Y)] = \frac{1}{c_w} \sum_x \sum_y p_\theta(y|x) \phi_{\ell}(x, y) \]

- **expected** probability that documents that contain *game* are labeled *hockey* (*c_w* is the count of *game*)

```
labels
baseball
hockey
politics
science
```

<table>
<thead>
<tr>
<th>contain game</th>
<th>(0.0 + 0.7 + 0.5 + 0.0 + 0.0) / 3 = 0.4</th>
</tr>
</thead>
</table>

(updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com))
Constraint Features & Expectations: Word Alignment

jugaban de una manera animada y muy cordial

it was an animated, very convivial game
Constraint Features & Expectations: Word Alignment

• \textbf{constraint feature:}

\[ \phi_m(x, y) = \# \text{ target words that align with } \textit{animada} \]
Constraint Features & Expectations:
Word Alignment

- **constraint feature:**
  \[ \phi_m(x, y) = \# \text{ target words that align with } \text{animada} \]

- **expectation:**
  \[ \mathbb{E}_{p_\theta}[\phi_m(x, y)] = \sum_y p_\theta(y|x) \phi_m(x, y) \]

- **expected** \# target words that align with *animada*
Constraint Features & Expectations: Word Alignment

- **constraint feature:**
  \( \phi_m(x, y) = \# \text{ target words that align with } \text{animada} \)

- **expectation:**
  \[ E_{p_\theta}[\phi_m(x, y)] = \sum_y p_\theta(y|x)\phi_m(x, y) \]

- **expected** \# target words that align with *animada*

```
  \[\begin{array}{ccc}
  \phi_1 = 1 & \phi_1 = 2 & \phi_1 = 1 \\
  \includegraphics[width=0.3\textwidth]{constraint1} & \includegraphics[width=0.3\textwidth]{constraint2} & \includegraphics[width=0.3\textwidth]{constraint3} \\
  p_\theta(y|x) = 0.7 & p_\theta(y|x) = 0.2 & p_\theta(y|x) = 0.1
  \end{array}\]
```
Constraint Features & Expectations: Word Alignment

- **constraint feature:**
  \[ \phi_m(x, y) = \# \text{ target words that align with } \textit{animada} \]

- **expectation:**
  \[ E_{p_\theta}[\phi_m(x, y)] = \sum_y p_\theta(y|x)\phi_m(x, y) \]

- **expected** \# target words that align with \textit{animada}

\[
\begin{align*}
\phi_1 = 1 & \quad \phi_1 = 2 \quad \phi_1 = 1 \\
\begin{array}{ccc}
\bigcirc & \bigcirc & \bigcirc \\
\bigcirc & \bigcirc & \bigcirc \\
\bigcirc & \bigcirc & \bigcirc \\
\end{array} & \begin{array}{ccc}
\bigcirc & \bigcirc & \bigcirc \\
\bigcirc & \bigcirc & \bigcirc \\
\bigcirc & \bigcirc & \bigcirc \\
\end{array} & \begin{array}{ccc}
\bigcirc & \bigcirc & \bigcirc \\
\bigcirc & \bigcirc & \bigcirc \\
\bigcirc & \bigcirc & \bigcirc \\
\end{array}
\end{align*}
\]

- \[ p_\theta(y|x) = 0.7 \]
- \[ p_\theta(y|x) = 0.2 \]
- \[ p_\theta(y|x) = 0.1 \]

\[ 0.7 \times 1 + \]

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
Constraint Features & Expectations: Word Alignment

• **constraint feature:**
  \[ \phi_m(x, y) = \# \text{ target words that align with } \text{animada} \]

• **expectation:**
  \[ \mathbb{E}_{p_\theta} [\phi_m(x, y)] = \sum_y p_\theta(y|x) \phi_m(x, y) \]

• **expected** \# target words that align with *animada*

\[
\begin{align*}
\phi_1 &= 1 \\
\phi_1 &= 2 \\
\phi_1 &= 1 \\
p_\theta(y|x) &= 0.7 \\
p_\theta(y|x) &= 0.2 \\
p_\theta(y|x) &= 0.1 \\
0.7 \times 1 &+ 0.2 \times 2 \\
&= 3.1
\end{align*}
\]
Constraint Features & Expectations: Word Alignment

- **constraint feature:**
  \[ \phi_m(x, y) = \# \text{ target words that align with animada} \]

- **expectation:**
  \[ \mathbb{E}_{p_\theta} [\phi_m(x, y)] = \sum_y p_\theta(y|x) \phi_m(x, y) \]

- **expected # target words that align with animada**

\[
\begin{align*}
\phi_1 &= 1 \\
\phi_1 &= 2 \\
\phi_1 &= 1
\end{align*}
\]

\[
\begin{align*}
p_\theta(y|x) &= 0.7 \\
p_\theta(y|x) &= 0.2 \\
p_\theta(y|x) &= 0.1
\end{align*}
\]

\[0.7 \times 1 + 0.2 \times 2 + 0.1 \times 1 = 1.2\]
Constraining Model Expectations
Constraining Model Expectations

- express preferences using target values: $b$
Constraining Model Expectations

• express preferences using **target values**: $b$

• **Example #1 Constraint**: $E_{p_\theta} [\phi_{w\ell}(X, Y)] \approx b$
  
  • *label distribution for game* is close to [40% 40% 20%]
Constraining Model Expectations

• express preferences using target values: $b$

• Example #1 Constraint: $E_{p_\theta} [\phi_w l(X, Y)] \approx b$
  • label distribution for game is close to [40% 40% 20%]

• Example #2 Constraint: $E_{p_\theta} [\phi_m (x, y)] \leq b$
  • expected number of target words that align with animada is at most 1
Preview: Document Classification
User Experiments [Druck et al. 08]

targets set with simple heuristic: majority label gets 90% of mass

complete set of labeled features

<table>
<thead>
<tr>
<th>PC</th>
<th>Mac</th>
</tr>
</thead>
<tbody>
<tr>
<td>dos</td>
<td>mac</td>
</tr>
<tr>
<td>ibm</td>
<td>apple</td>
</tr>
<tr>
<td>hp</td>
<td>quadra</td>
</tr>
<tr>
<td>dx</td>
<td></td>
</tr>
</tbody>
</table>

updated slides: http://sideinfo.wikkii.com
Preview: Document Classification
User Experiments [Druck et al. 08]

PC vs. Mac

~15 minutes, 100 documents labeled (or skipped): 78% accuracy

Targets set with simple heuristic: majority label gets 90% of mass

Complete set of labeled features

<table>
<thead>
<tr>
<th>PC</th>
<th>Mac</th>
</tr>
</thead>
<tbody>
<tr>
<td>dos</td>
<td>mac</td>
</tr>
<tr>
<td>ibm</td>
<td>apple</td>
</tr>
<tr>
<td>hp</td>
<td>quadra</td>
</tr>
<tr>
<td>dx</td>
<td></td>
</tr>
</tbody>
</table>

updated slides: http://sideinfo.wikii.com
Preview: Document Classification
User Experiments [Druck et al. 08]

~2 minutes, 100 features labeled (or skipped):
~15 minutes, 100 documents labeled (or skipped):
82% accuracy 78% accuracy

PC vs. Mac

targets set with simple heuristic: majority label gets 90% of mass

complete set of labeled features

<table>
<thead>
<tr>
<th>PC</th>
<th>Mac</th>
</tr>
</thead>
<tbody>
<tr>
<td>dos</td>
<td>mac</td>
</tr>
<tr>
<td>ibm</td>
<td>apple</td>
</tr>
<tr>
<td>hp</td>
<td>quadra</td>
</tr>
<tr>
<td>dx</td>
<td></td>
</tr>
</tbody>
</table>

updated slides: http://sideinfo.wikii.com
Preview: Word Alignment
[Graça et al. 10]

- HMM
- HMM + Bijectivity

updated slides: http://sideinfo.wikii.com
Preview: Word Alignment

[Graça et al. 10]

<table>
<thead>
<tr>
<th>Precision</th>
<th>HMM</th>
<th>HMM + Bijectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>86.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>77.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>68.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60</td>
<td>En-Pt</td>
<td>Pt-En</td>
</tr>
</tbody>
</table>

updated slides: http://sideinfo.wikii.com
Preview: Word Alignment

[Graça et al. 10]

- En-Pt
- Pt-En
- En-Es
- Es-En

- Precision
  - HMM
  - HMM + Bijectivity

- 95
- 86.25
- 77.5
- 68.75
- 60

updated slides: http://sideinfo.wikii.com
Preview: Word Alignment

[Graça et al. 10]

Updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Preview: Word Alignment

[Graca et al. 10]

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>HMM</th>
<th>HMM + Bijectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-Pt</td>
<td>68.75</td>
<td>77.5</td>
</tr>
<tr>
<td>Pt-En</td>
<td>68.75</td>
<td></td>
</tr>
<tr>
<td>En-Es</td>
<td>95</td>
<td>86.25</td>
</tr>
<tr>
<td>Es-En</td>
<td>60</td>
<td></td>
</tr>
</tbody>
</table>
Preview: Word Alignment

[Graça et al. 10]
Preview: Word Alignment

[Graça et al. 10]

updated slides: http://sideinfo.wikii.com
Preview: Word Alignment
[Graca et al. 10]

- En-Pt
- Pt-En
- En-Es
- Es-En

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>HMM</th>
<th>HMM + Bijectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>60</td>
<td>77.5</td>
</tr>
<tr>
<td>En-Pt</td>
<td>86.25</td>
<td></td>
</tr>
<tr>
<td>Pt-En</td>
<td>68.75</td>
<td></td>
</tr>
<tr>
<td>En-Es</td>
<td>66.75</td>
<td></td>
</tr>
<tr>
<td>Es-En</td>
<td>66.75</td>
<td></td>
</tr>
</tbody>
</table>

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Preview: Word Alignment

[Graca et al. 10]

HMM

HMM + Bijectivity

updated slides: http://sideinfo.wikii.com
Preview: Word Alignment

[Graça et al. 10]

updated slides: http://sideinfo.wikii.com
Preview: Word Alignment
[Graca et al. 10]

- En-Pt
- Pt-En
- En-Es
- Es-En

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>HMM</th>
<th>HMM + Bijectivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-Pt</td>
<td>60</td>
<td>68.75</td>
</tr>
<tr>
<td>Pt-En</td>
<td>60</td>
<td>68.75</td>
</tr>
<tr>
<td>En-Es</td>
<td>77.5</td>
<td>86.25</td>
</tr>
<tr>
<td>Es-En</td>
<td>77.5</td>
<td>95</td>
</tr>
</tbody>
</table>

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Some related frameworks

updated slides: http://sideinfo.wikki.com
Some related frameworks

updated slides: http://sideinfo.wikii.com
For concreteness: running example

Want to ensure that 25% of unlabeled documents are about politics

- **constraint** features

  \[ \phi(x, y) = \begin{cases} 
  1 & \text{if } y \text{ is “politics”} \\
  0 & \text{otherwise} 
  \end{cases} \]

- preferred expected value

  \[ b = 0.25 \]

- Expectation w.r.t. unlabeled data
Constraint-Driven Learning

University of Illinois at Urbana-Champaign (2007)

Application: Information Extraction

Idea: Tell the system:
• Citations have contiguous authors
• Citation fields usually end with punctuation

Implementation:
• Design a penalty function to encode constraint
Constraint-Driven Learning

University of Illinois at Urbana-Champaign (2007)

Idea: Use knowledge to decode better:
predict 25% of articles are “politics”

\[
\hat{Y} = \arg \max_Y \log p_\theta(Y|X) - \text{penalty}(Y)
\]
Constraint-Driven Learning

University of Illinois at Urbana-Champaign (2007)

Idea: Use knowledge to decode better: predict 25% of articles are “politics”

Idea: Retrain with predictions.

Constraint Driven Learning:

E-Step: set $\hat{Y} = \arg \max_Y \log p_\theta(Y|X) - \text{penalty}(Y)$

M-Step: set $\theta = \arg \max_\theta \log p_\theta(\hat{Y}|X)$

updated slides: http://sideinfo.wikii.com
Constraint-Driven Learning

**Motivation:** Hard EM-like algorithm with preferences

**Constraint Driven Learning:**

E-Step: set $\hat{Y} = \arg\max_Y \log p_\theta(Y|X) - \text{penalty}(Y)$

M-Step: set $\theta = \arg\max_\theta \log p_\theta(\hat{Y}|X)$

- penalties encode similar information as $E[\phi] \approx b$

$$\text{penalty}(Y) = ||\phi(X, Y) - b||_\beta$$

- E-Step can be hard; use beam search

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
Generalized Expectation Constraints


University of Massachusetts Amherst (2007)

Application: Document Classification, Info Extraction

Idea: Use labeled features:

- Document has “puck” $\Rightarrow p(\text{class} = \text{sport}) = 90\%$

Implementation:

- Add penalty while training:

$$\max_\theta \mathcal{L}_\theta \Rightarrow \max_\theta \mathcal{L}_\theta + \text{penalty}(p_\theta(Y|X))$$

updated slides: http://sideinfo.wikii.com
Generalized Expectation Constraints

University of Massachusetts Amherst (2007)

**Idea:** Penalize “bad” distributions:
train a model to predict 25% of articles as “politics”
Generalized Expectation Constraints


University of Massachusetts Amherst (2007)

Idea: Penalize “bad” distributions:
train a model to predict 25% of articles as “politics”

Objective:

\[
\max_{\theta} {\mathcal L} (\theta; D_L) \quad \text{where} \\
E_{p^\theta}(Y|X)[\phi] = E_{p^\theta}(Y|X)[\phi(X, Y)] \\
= \sum_Y p^\theta(Y|X) \phi(X, Y) \text{ is short-hand}
\]

Optimization: gradient descent on \(\theta\)

updated slides: http://sideinfo.wikki.com
Posterior Regularization

University of Pennsylvania (2007)

**Application:** Word alignment for machine translation

**Idea:** Ensure reasonable alignments during training:
- Bijectivity: each word aligns to at most one word
- Symmetry: $\text{En} \rightarrow \text{Fr}$ and $\text{Fr} \rightarrow \text{En}$ give same alignment

**Implementation:** EM algorithm with “valid” distribution.

$E'$-Step: set $q(Y) = \arg\min \quad \mathcal{D}_{KL}(q(Y)\|p_\theta(Y|X))$

sane distribution $q$

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Posterior Regularization
University of Pennsylvania (2007)

Idea: EM algorithm with valid posteriors

Define: Valid posteriors: \( Q = \{ q(\mathbf{Y}) : E_q[\phi] \approx b \} \)
e.g. \( q(\mathbf{Y}) \) that assign 25% articles to “politics”

EM:
E-Step: set \( q(\mathbf{Y}) = p_\theta(\mathbf{Y}|\mathbf{X}) \)

Valid posteriors:
E-Step: set \( q(\mathbf{Y}) = \arg \min_{q \in Q} D_{KL}(q(\mathbf{Y})||p_\theta(\mathbf{y}|\mathbf{x})) \)
Posterior Regularization


University of Pennsylvania (2007)

Idea: EM algorithm with valid posteriors

Define: Valid posteriors: $Q = \{ q(Y) : \mathbb{E}_q[\phi] \approx b \}$
e.g: $q(Y)$ that assign 25% articles to “politics”

EM:

E-Step: set $q(Y) = p_\theta(Y|X)$
M-Step: set $\theta = \arg \max_\theta \mathbb{E}_{q(Y)}[p_\theta(Y|X)]$

Constrained EM:

E-Step: set $q(Y) = \arg \min_{q \in Q} \mathcal{D}_{KL}(q(Y)\|p_\theta(y|x))$
M-Step: set $\theta = \arg \max_\theta \mathbb{E}_{q(Y)}[p_\theta(Y|X)]$

updated slides: http://sideinfo.wikki.com
Posterior Regularization

**Idea:** define $Q$: set of $q$ such that $E_q[\phi] \approx b$

**Constrained EM:**

E-Step: set $q(Y) = \arg\min_{q \in Q} D_{KL}(q(Y) \| p_\theta(y|x))$

M-Step: set $\theta = \arg\max_\theta E_{q(Y)}[p_\theta(Y|X)]$

**Objective:**

$$\max_\theta \mathcal{L}(\theta) - D_{KL}(Q \| p_\theta(Y|X))$$

where

$D_{KL}(q\|p) = E_q \left[ \log \frac{q}{p} \right]$ is Kullback-Leibler divergence

$D_{KL}(Q\|p) = \min_{q \in Q} D_{KL}(q\|p)$

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
Posterior Regularization

**Hard constraints:**

\[
\max \limits_\theta \mathcal{L}(\theta) - \min \limits_{q \in \mathcal{Q}} \mathcal{D}_{KL}(q(Y)\| p_\theta(Y|X))
\]

\[
\mathcal{Q} = \left\{ q(Y) : \|E_q[\phi(Y)] - b\|_2^2 \leq \epsilon \right\}
\]

**Soft constraints:**

\[
\max \limits_\theta \mathcal{L}(\theta) - \min \limits_{q} \left( \mathcal{D}_{KL}(q(Y)\| p_\theta(Y|X)) + \alpha \|E_q[\phi(Y)] - b\|_2^2 \right)
\]
Summary: CoDL, GE, PR

Constraint Driven Learning:
Apply constraints at decode time + self-training.

\[
\arg\max_Y \log p_\theta(Y|X) - \text{penalty}(Y)
\]

Generalized Expectation Constraints:
Train model to satisfy constraints.

\[
\max_\theta \mathcal{L}_\theta \implies \max_\theta \mathcal{L}_\theta - \text{penalty}(p_\theta(Y|X))
\]

Posterior Regularization:
Project onto a constraint set + EM training.

\[
\max_\theta \mathcal{L}_\theta \implies \max_\theta \mathcal{L}(\theta; D_L) - \mathcal{D}_{KL}(Q|| p_\theta(Y|X))
\]

updated slides: http://sideinfo.wikikii.com
A Bayesian View: Measurements

P. Liang, M. Jordan, D. Klein (2009)
University of California, Berkeley (2009)

**Idea:** Bayesian formulation for learning with constraints:
- Nature computes the hidden value: $\phi(X, Y)$
- We observe $b = \phi(X, Y) + \text{noise}$

**Bonus:** Relates the frameworks above.

**Figure:** The model used by Liang et al. using our notation. We have separated some noisy version of the labeled corpus

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
**A Bayesian View: Measurements**

P. Liang, M. Jordan, D. Klein (2009)

**Objective:** mode of $\theta$ given observations

$$
\max_{\theta} \log p(\theta) + \sum_{(x,y) \in D_L} \log p_{\theta}(y|x) = \mathcal{L}(\theta; X_L, Y_L)
$$
A Bayesian View: Measurements

P. Liang, M. Jordan, D. Klein (2009)

Objective: mode of $\theta$ given observations

$$\max_{\theta} \mathcal{L}(\theta; X_L, Y_L)$$
A Bayesian View: Measurements

P. Liang, M. Jordan, D. Klein (2009)

\[
\begin{align*}
X_L & \quad \theta & \quad X \\
Y_L & \quad & \\
\end{align*}
\]

**Objective:** mode of \( \theta \) given observations

\[
\max_{\theta} \mathcal{L}(\theta; X_L, Y_L)
\]
A Bayesian View: Measurements

P. Liang, M. Jordan, D. Klein (2009)

Objective: mode of $\theta$ given observations

$$\max_{\theta} \mathcal{L}(\theta; X_L, Y_L)$$
A Bayesian View: Measurements

P. Liang, M. Jordan, D. Klein (2009)

Objective: mode of $\theta$ given observations

$$\max_{\theta} \mathcal{L}(\theta; X_L, Y_L)$$

updated slides: http://sideinfo.wikiki.com
A Bayesian View: Measurements

P. Liang, M. Jordan, D. Klein (2009)

\[ \text{Objective: mode of } \theta \text{ given observations} \]

\[
\max_{\theta} \quad \mathcal{L}(\theta; X_L, Y_L) + \log \mathbb{E}_{p_\theta(Y|X)} [p_N(b|\phi(X, Y))] 
\]

where \( p_N(b|\phi(X, Y)) \) models the noise in observing \( b \)

updated slides: http://sideinfo.wikii.com
Objective: mode of $\theta$ given observations

$$\max_{\theta} \mathcal{L}(\theta; X_L, Y_L) + \log \mathbb{E}_{p_{\theta}(Y|X)} [p_N(b|\phi(X, Y))]$$

where $p_N(b|\phi(X, Y))$ models the noise in observing $b$

Great! How do I optimize this?

updated slides: http://sideinfo.wikki.com
What's wrong with this picture?

**Objective:** mode of $\theta$ given observations

$$\max_\theta \quad \mathcal{L}(\theta; X_L, Y_L) + \log \mathbb{E}_{p_\theta(Y|X)} \left[ p_N(b|\phi(X, Y)) \right]$$

**Example:** Noise free: exactly 25% of articles are “politics”

$$p_N(b|\phi(X, Y)) = \begin{cases} 1 \quad \text{if } b = \phi(X, Y) \\ 0 \quad \text{otherwise} \end{cases}$$
What's wrong with this picture?

**Objective:** mode of $\theta$ given observations

$$\max_{\theta} \mathcal{L}(\theta; X_L, Y_L) + \log \mathbf{E}_{p_\theta}(Y|X) \left[ p_N(b|\phi(X, Y)) \right]$$

**Example:** Noise free: exactly 25% of articles are “politics”

$$p_N(b|\phi(X, Y)) = \begin{cases} 1 & \text{if } b = \phi(X, Y) \\ 0 & \text{otherwise} \end{cases} = 1(b = \phi)$$

What is the probability exactly 25% of the articles are labeled ``politics"?
What's wrong with this picture?

Objective: mode of $\theta$ given observations

$$\max_{\theta} \ L(\theta; X_L, Y_L) + \log E_{p\theta}(Y|X) \left[ p_N(b|\phi(X, Y)) \right]$$

Example: Noise free: exactly 25% of articles are “politics”

$$p_N(b|\phi(X, Y)) = \begin{cases} 1 & \text{if } b = \phi(X, Y) \\ 0 & \text{otherwise} \end{cases} = 1(b = \phi)$$

What is the probability exactly 25% of the articles are labeled ``politics''?

$$E_{p\theta}(Y|X) \left[ 1(b = \phi(X, Y)) \right]$$

How do we optimize this with respect to $\theta$?
What's wrong with this picture?
What's wrong with this picture?

**Example:** Compute prob: 25% of docs are “politics”.

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
What's wrong with this picture?

**Example:** Compute prob: 25% of docs are “politics”.

<table>
<thead>
<tr>
<th>Article</th>
<th>$p(\text{“politics”})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
</tr>
</tbody>
</table>
What's wrong with this picture?

**Example:** Compute prob: 25% of docs are “politics”.

<table>
<thead>
<tr>
<th>Article</th>
<th>p(“politics”)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Naively:

\[
0.2 \times (1 - 0.4) \times (1 - 0.1) \times (1 - 0.6) \\
+ \ldots + \\
+(1 - 0.2) \times (1 - 0.4) \times (1 - 0.1) \times 0.6
\]
What's wrong with this picture?

Example: Compute prob: 25% of docs are “politics”.

<table>
<thead>
<tr>
<th>Article</th>
<th>p(“politics”)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Naively:

\[
0.2 \times (1 - 0.4) \times (1 - 0.1) \times (1 - 0.6) \\
+ \ldots + \\
+(1 - 0.2) \times (1 - 0.4) \times (1 - 0.1) \times 0.6
\]

For one constraint, maybe we can make a specialized routine. If there are many constraints, that doesn’t work.
What's wrong with this picture?

**Example:** Compute prob: 25% of docs are “politics”.

<table>
<thead>
<tr>
<th>Article</th>
<th>p(“politics”)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Naively:

\[
0.2 \times (1 - 0.4) \times (1 - 0.1) \times (1 - 0.6) + \ldots + (1 - 0.2) \times (1 - 0.4) \times (1 - 0.1) \times 0.6
\]

For one constraint, maybe we can make a specialized routine. If there are many constraints, that doesn’t work.

**Easier:** What is the expected number of “politics” articles?

\[
0.2 + 0.4 + 0.1 + 0.6
\]

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Probabilities and Expectations
Probabilities and Expectations

**Easier:** Compute expected number of “politics” docs.
Probabilities and Expectations

**Easier:** Compute expected number of “politics” docs.

\[ \phi(X, Y) = \text{count } \# \text{ of “politics” docs} \]
Probabilities and Expectations

Easier: Compute expected number of “politics” docs.

\[ \phi(X, Y) = \text{count } \# \text{ of “politics” docs} \]

\[ \phi(X, Y) = \sum_{y_i \in Y} \phi(y_i) \]

\[ \phi(y_i) = \begin{cases} 
1 & \text{if } y_i = \text{“politics”} \\
0 & \text{otherwise} 
\end{cases} \]
Probabilities and Expectations

**Easier:** Compute expected number of “politics” docs.

\[ \phi(X, Y) = \text{count } \# \text{ of “politics” docs} \]

\[ \phi(X, Y) = \sum_{y_i \in Y} \phi(y_i) \]

\[ \phi(y_i) = \begin{cases} 
1 & \text{if } y_i = \text{“politics”} \\
0 & \text{otherwise} 
\end{cases} \]

\[ E[\phi(X, Y)] = E \left[ \sum_{y_i \in Y} \phi(y_i) \right] \]
Probabilities and Expectations

**Easier:** Compute expected number of “politics” docs.

\[ \phi(X, Y) = \text{count } \# \text{ of “politics” docs} \]

\[ \phi(X, Y) = \sum_{y_i \in Y} \phi(y_i) \]

\[ \phi(y_i) = \begin{cases} 
1 & \text{if } y_i = \text{“politics”} \\
0 & \text{otherwise} 
\end{cases} \]

\[
E[\phi(X, Y)] = E \left[ \sum_{y_i \in Y} \phi(y_i) \right] = \sum_i E[\phi(y_i)]
\]

(by linearity of expectations)

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Probabilities and Expectations

**Hard:** Compute probability 25% of docs are “politics”.

**Easy:** Compute expected number of “politics” docs.

\[
\text{Article} & \quad p(\text{“politics”}) \\
1 & 0.2 \\
2 & 0.4 \\
3 & 0.1 \\
4 & 0.6
\]

\[
= 0.2 + 0.4 + 0.1 + 0.6 = 1.3
\]
Probabilities and Expectations

**Hard:** Compute probability 25% of docs are “politics”.

**Easy:** Compute expected number of “politics” docs.

$E[\phi(X, Y)] = E \left[ \sum_{y_i \in Y} \phi(y_i) \right] = \sum_i E[\phi(y_i)]$

$= 0.2 + 0.4 + 0.1 + 0.6 = 1.3$

<table>
<thead>
<tr>
<th>Article</th>
<th>p(“politics”)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
</tr>
<tr>
<td>2</td>
<td>0.4</td>
</tr>
<tr>
<td>3</td>
<td>0.1</td>
</tr>
<tr>
<td>4</td>
<td>0.6</td>
</tr>
</tbody>
</table>
More Generally

\[ p_\theta(X, Y) = \prod_{c} \psi_\theta(X, y_c) \]

\[ p_\theta(X, y_c) \]

E.g. forward-backward, inside-outside

updated slides: http://sideinfo.wikii.com 66
More Generally

**Observation:** for many models in NLP:

\[ p_\theta(X, Y) = \prod_c \psi_\theta(X, y_c) \]

and it’s easy to compute: \( p_\theta(X, y_c) \)

E.g. forward-backward, inside-outside
More Generally

**Observation:** for many models in NLP:

\[ p_\theta(X, Y) = \prod_c \psi_\theta(X, y_c) \]

and it’s easy to compute: \( p_\theta(X, y_c) \)

so if: \( \phi(X, Y) = \sum_c \phi(X, y_c) \)

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
More Generally

**Observation:** for many models in NLP:

\[ p_\theta(X, Y) = \prod_c \psi_\theta(X, y_c) \]

and it’s easy to compute: \( p_\theta(X, y_c) \)

so if: \( \phi(X, Y) = \sum_c \phi(X, y_c) \)

we can compute: \( E[\phi(X, Y)] = \sum_c E[\phi(X, y_c)] \)

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
More Generally

**Idea:** Approximate:

\[
E_{p_\theta(Y|X)} \left[ p_N \left( b \mid \phi(X, Y) \right) \right] \approx p_N \left( b \mid E_{p_\theta(Y|X)} \left[ \phi(X, Y) \right] \right)
\]

**Example:** Gaussian noise:

\[
p_N \left( b \mid E[\phi] \right) = \frac{1}{Z_N} \exp \left( -\frac{||b - E[\phi]||^2}{2\sigma^2} \right)
\]

updated slides: http://sideinfo.wikii.com
Probabilities and Expectations
Probabilities and Expectations

**Approximation:** \( E_{p_\theta(Y|X)} \left[ p_N \left( b \mid \phi \right) \right] \approx p_N \left( b \mid E_{p_\theta(Y|X)} [\phi] \right) \)

\[ \Downarrow \]

**Objective:** \( \max_\theta \mathcal{L}(\theta; X_L, Y_L) + \log p_N \left( b \mid E_{p_\theta(Y|X)} [\phi] \right) \)
Probabilities and Expectations

**Approximation:** \( \mathbb{E}_{\theta}(Y|X) \left[ p_{N} \left( b \mid \phi \right) \right] \approx p_{N} \left( b \mid \mathbb{E}_{\theta}(Y|X) \left[ \phi \right] \right) \)

\[ \downarrow \]

**Objective:** \( \max_{\theta} \mathcal{L}(\theta; X_{L}, Y_{L}) + \log p_{N} \left( b \mid \mathbb{E}_{\theta}(Y|X) \left[ \phi \right] \right) \)

**Example:** Gaussian noise:

\[ p_{N} \left( b \mid \mathbb{E} \left[ \phi \right] \right) = \frac{1}{Z_{N}} \exp \left( -\frac{||b - \mathbb{E}[\phi]||_{2}^{2}}{2\sigma^{2}} \right) \]
Probabilities and Expectations

Approximation: \( \mathbb{E}_{p_\theta(Y|X)} \left[ p_N \left( b \mid \phi \right) \right] \approx p_N \left( b \mid \mathbb{E}_{p_\theta(Y|X)} [\phi] \right) \)

Objective: \[ \max_{\theta} \mathcal{L}(\theta; X_L, Y_L) + \log p_N \left( b \mid \mathbb{E}_{p_\theta(Y|X)} [\phi] \right) \]

Example: Gaussian noise:

\[ p_N \left( b \mid \mathbb{E} [\phi] \right) = \frac{1}{Z_N} \exp \left( - \frac{\| b - \mathbb{E}[\phi] \|_2^2}{2\sigma^2} \right) \]

\[ \log p_N \left( b \mid \mathbb{E} [\phi] \right) \Rightarrow - \| \mathbb{E}[\phi] - b \|_2^2 \]

so for appropriate \( \log p_N \left( b \mid \mathbb{E}[\phi] \right) \) this is identical to GE!

updated slides: http://sideinfo.wikki.com

68
Computing and Optimizing GE

\[
\max_\theta \mathcal{L}(\theta) - \left\| \mathbb{E}_{p_\theta(Y|X)}[\phi] - b \right\|_\beta
\]

- Easy to compute objective if: \( \phi(Y, X) \) decomposes.

updated slides: http://sideinfo.wikii.com
Computing and Optimizing GE

**GE Optimization**: Gradient ascent (or L-BFGS)

$$\max_{\theta} \mathcal{L}(\theta) - \left\| \mathbb{E}_{p_{\theta}(Y|X)}[\phi] - b \right\|_\beta$$

- Easy to compute objective if: $\phi(Y, X)$ decomposes.

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Computing and Optimizing GE

**GE Optimization**: Gradient ascent (or L-BFGS)

$$\max_\theta \mathcal{L}(\theta) - \| \mathbb{E}_{p_{\theta}(Y|X)}[\phi] - b \|_\beta$$

- Gradient computation:
Computing and Optimizing GE

**GE Optimization:** Gradient ascent (or L-BFGS)

\[
\max_\theta \mathcal{L}(\theta) - \| \mathbb{E}_{p_{\theta}(Y|X)}[\phi] - b \|_\beta
\]

- Gradient computation:

\[
\frac{\partial}{\partial \theta} \| \mathbb{E}_{p}[\phi] - b \|^2_2 = (\mathbb{E}[\phi] - b) \frac{\partial}{\partial \theta} \mathbb{E}_{p}[\phi]
\]
Computing and Optimizing GE

**GE Optimization:** Gradient ascent (or L-BFGS)

\[
\max_\theta \mathcal{L}(\theta) - \| \mathbf{E}_{p(\mathbf{Y}|\mathbf{X})}[\phi] - \mathbf{b} \|_\beta
\]

- Gradient computation:

\[
\frac{\partial}{\partial \theta} \| \mathbf{E}_p[\phi] - \mathbf{b} \|^2_2 = (\mathbf{E}[\phi] - \mathbf{b}) \frac{\partial}{\partial \theta} \mathbf{E}_p[\phi]
\]

\[p_\theta(\mathbf{Y}|\mathbf{X}) \propto \exp(\theta \cdot f(\mathbf{X}, \mathbf{Y}))\]
Computing and Optimizing GE

**GE Optimization:** Gradient ascent (or L-BFGS)

$$\max_{\theta} L(\theta) - \| E_{p_{\theta}(Y|X)}[\phi] - b \|_\beta$$

- Gradient computation:

$$\frac{\partial}{\partial \theta} \| E_p[\phi] - b \|^2_2 = (E[\phi] - b) \frac{\partial}{\partial \theta} E_p[\phi]$$

$$p_{\theta}(Y|X) \propto \exp(\theta \cdot f(X, Y))$$

$$\frac{\partial}{\partial \theta} E_p[\phi] = E[\phi] \times E[f] - E[\phi \times f]$$

Computing $E[\phi \times f]$ can be hard sometimes.
Computing and Optimizing GE

GE Objective:

\[ \max_{\theta} \mathcal{L}(\theta) - \left\| \mathbb{E}_{p_{\theta}(Y|X)}[\phi] - b \right\|_\beta \]

Gradient involves \( \mathbb{E}[\phi \times f] \)

Gradient computation:

\[ \mathbb{E}[\phi \times f] = \sum_Y p(Y) \phi(Y) \times f(Y) \]
Computing and Optimizing GE

**GE Objective:**

\[
\max_{\theta} \mathcal{L}(\theta) - \left\| \mathbb{E}_{p_{\theta}(Y|X)}[\phi] - b \right\|_\beta
\]

Gradient involves \( \mathbb{E}[\phi \times f] \)

Gradient computation:

\[
\mathbb{E}[\phi \times f] = \sum_Y p(Y) \phi(Y) \times f(Y)
\]

\[
\phi(Y) \times f(Y) = \left[ \sum_i \phi(y_i) \right] \times \left[ \sum_j f(y_j) \right]
\]
Computing and Optimizing GE

**GE Objective:**

$$\max_{\theta} L(\theta) - \| E_{p_\theta(Y|X)}[\phi] - b \|_\beta$$

Gradient involves $E[\phi \times f]$

Gradient computation:

$$E[\phi \times f] = \sum_Y p(Y) \phi(Y) \times f(Y)$$

$$\phi(Y) \times f(Y) = \left[ \sum_i \phi(y_i) \right] \times \left[ \sum_j f(y_j) \right]$$

$$= \ldots + \phi(y_i) \times f(y_j) + \ldots$$

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
Example dynamic program

\[
\mathbb{E}_{p_\theta(Y|X)}[f(Y)] = \sum_{y_i} \mathbb{E}_{p_\theta(Y|X)}[f(y_i)]
\]
Example dynamic program

\[ E_{p\theta}(Y|X)[f(Y)] = \sum_{y_i} E_{p\theta}(Y|X)[f(y_i)] \]

Just need very local information for \( E_{p\theta}(Y|X)[f(y_i)] \)

updated slides: http://sideinfo.wikki.com
Optimizing GE Objective

\[
E_{p\theta(Y|X)}[f(Y) \times \phi(Y)] = \sum_{y_i} \sum_{y_j} E_{p\theta(Y|X)}[f(y_i) \times \phi(y_j)]
\]
Optimizing GE Objective

$$\mathbb{E}_{p_{\theta}(Y|X)}[f(Y) \times \phi(Y)] = \sum_{y_i} \sum_{y_j} \mathbb{E}_{p_{\theta}(Y|X)}[f(y_i) \times \phi(y_j)]$$

Need a modified dynamic program for computing

$$\mathbb{E}_{p_{\theta}(Y|X)}[f(y_i) \times \phi(y_j)]$$

updated slides: http://sideinfo.wikki.com
A Variational Approximation

\[ O_{GE} = \max_{\theta} \mathcal{L}(\theta) - \| b - E_{p\theta} [\phi] \|_{\beta} \]

\[ \frac{\partial}{\partial \theta} \| b - E_{p\theta} [\phi] \|_{\beta} \]
A Variational Approximation

GE Objective:

\[ \mathcal{O}_{GE} = \max_{\theta} \mathcal{L}(\theta) - \| b - \mathbf{E}_{p_{\theta}}[\phi] \|_{\beta} \]

- Computing \( \frac{\partial}{\partial \theta} \| b - \mathbf{E}_{p_{\theta}}[\phi] \|_{\beta} \) can be hard.
A Variational Approximation

**GE Objective:**

\[ O_{GE} = \max_{\theta} \mathcal{L}(\theta) - \| b - \mathbb{E}_{p_{\theta}}[\phi] \|_\beta \]

- Computing \( \frac{\partial}{\partial \theta} \| b - \mathbb{E}_{p_{\theta}}[\phi] \|_\beta \) can be hard.

**Idea:** use variational approximation \( q(Y) \approx p_{\theta}(Y|X) \)

updated slides: [http://sideinfo.wikiki.com](http://sideinfo.wikiki.com)
A Variational Approximation

GE Objective:

\[ \mathcal{O}_{GE} = \max_{\theta} \mathcal{L}(\theta) - \| b - \mathbb{E}_{p_\theta} [\phi] \|_\beta \]

- Computing \( \frac{\partial}{\partial \theta} \| b - \mathbb{E}_{p_\theta} [\phi] \|_\beta \) can be hard.

Idea: use variational approximation \( q(Y) \approx p_\theta(Y|X) \)

Tie together \( \min_{q(Y)} \mathcal{D}_{KL} (q(Y) \| p_\theta(Y|X)) \)
A Variational Approximation

GE Objective:

\[ O_{GE} = \max_{\theta} \mathcal{L}(\theta) - \| b - E_{p_\theta}[\phi] \|_\beta \]

- Computing \( \frac{\partial}{\partial \theta} \| b - E_{p_\theta}[\phi] \|_\beta \) can be hard.

Idea: use variational approximation \( q(Y) \approx p_\theta(Y|X) \)

Tie together \( \min_{q(Y)} D_{KL}(q(Y) \| p_\theta(Y|X)) \)

\[ \max_{\theta, q(Y)} \mathcal{L}(\theta) - D_{KL}(q(Y) \| p_\theta(Y|X)) - \| E_q[\phi(X, Y)] - b \|_\beta \]

Benefit: \[ \frac{\partial}{\partial \theta} \| b - E_q[\phi] \|_\beta = 0 \]
A Variational Approximation

GE Objective:

\[ \mathcal{O}_{GE} = \max_\theta \mathcal{L}(\theta) - \| b - \mathbb{E}_{p_\theta}[\phi] \|_\beta \]

- Computing \( \frac{\partial}{\partial \theta} \| b - \mathbb{E}_{p_\theta}[\phi] \|_\beta \) can be hard.

Idea: use variational approximation \( q(Y) \approx p_\theta(Y|X) \)

Tie together \( \min_{q(Y)} \mathcal{D}_{KL} (q(Y) \| p_\theta(Y|X)) \)

\[ \max_{\theta, q(Y)} \mathcal{L}(\theta) - \mathcal{D}_{KL} (q(Y) \| p_\theta(Y|X)) - \| \mathbb{E}_q[\phi(X, Y)] - b \|_\beta \]

This is the PR objective!
Types of constraints

\[
\min_{q} \mathcal{D}_{KL}(q(Y) \| p_{\theta}(Y|X)) \quad \text{s.t.} \quad \| E_q[\phi] - b \|_\beta \leq \epsilon
\]

\[\phi(Y, X)\]
Types of constraints

**Posterior Regularization**: KL projection

\[
\min_q \mathcal{D}_{KL}(q(Y) \parallel p_{\theta}(Y|X)) \quad \text{s.t.} \quad \|E_q[\phi] - b\|_{\beta} \leq \epsilon
\]

Similar to a small maximum entropy problem

Optimize via gradient of dual \( \phi(Y, X) \)
Types of constraints

**Posterior Regularization:** KL projection

\[
\min_q D_{KL}(q(Y)\|p_\theta(Y|X)) \text{ s.t. } \|E_q[\phi] - b\|_\beta \leq \epsilon
\]

Similar to a small maximum entropy problem
Optimize via gradient of dual \( \phi(Y, X) \)
Need to compute \( E_q[\phi(X, Y)] \)

usually easy if \( \phi(X, Y) \) decomposes
Types of constraints

**Posterior Regularization:** KL projection

\[
\min_q \mathcal{D}_{KL}(q(Y) \| p_\theta(Y|X)) \quad \text{s.t.} \quad \| E_q[\phi] - b \|_\beta \leq \epsilon
\]

Similar to a small maximum entropy problem

Optimize via gradient of dual \( \phi(Y, X) \)

Need to compute \( E_q[\phi(X, Y)] \)

usually easy if \( \phi(X, Y) \) decomposes

**Otherwise:** Sample (e.g. K. Bellare, G. Druck, and A. McCallum, 2009)
Approximating with the mode

**PR Objective:**

\[
\max_{\theta, q(Y)} \mathcal{L}(\theta) - D_{KL}(q(Y) \parallel p_{\theta}(Y|X)) - \| E_q[\phi(X, Y)] - b \|_\beta
\]

What if we can’t hold \( q(Y) \) in memory?
Or we can’t compute expectations?
Or min-KL is hard?

**Idea:** use hard assignment \( q(Y) \approx 1(Y = \hat{Y}) \):
Approximating with the mode

**Idea:** use hard assignment \( q(Y) \approx 1(Y = \hat{Y}) \):

**KL-projection:**

\[
\min_{q(Y)} D_{KL} (q(Y) \parallel p_\theta(Y|X)) + \| E_q[\phi(X, Y)] - b \|_\beta
\]

\[
D_{KL} (q(Y) \parallel p_\theta(Y|X)) = \sum_Y q(Y) \log \frac{q(Y)}{p_\theta(Y|X)}
\]

\[
\Rightarrow - \log p_\theta(\hat{Y}|X)
\]

\[
\| E_q[\phi(X, Y)] - b \|_\beta \Rightarrow \| \phi(X, \hat{Y}) - b \|_\beta
\]

\[
= - \log p_N(b|\phi(X, \hat{Y}))
\]
Approximating with the mode

Idea: use hard assignment \( q(Y) \approx 1(Y = \hat{Y}) \):

KL-projection:
\[
\min_{q(Y)} \mathcal{D}_{KL}(q(Y) \parallel p_\theta(Y|X)) + \|E_q[\phi(X, Y)] - b\|_\beta
\]
\[
\Downarrow
\]
\[
\max_Y \log(p_\theta(Y)) + \log p_N(b|\phi(X, Y))
\]

Use the normal M-step for hard-EM.

updated slides: http://sideinfo.wikii.com
Approximating with the mode

Idea: use hard assignment \( q(Y) \approx 1(Y = \hat{Y}) \):

KL-projection:
\[
\min_{q(Y)} D_{KL}(q(Y) \parallel p_\theta(Y|X)) + \| E_q[\phi(X, Y)] - b \|_\beta
\]

\[
\max_Y \log(p_\theta(Y)) + \log p_N(b|\phi(X, Y)) = -\text{penalty}(\hat{Y})
\]

Use the normal M-step for hard-EM.

This is the CoDL algorithm!
Approximating with the mode

**Idea:** use hard assignment \( q(Y) \approx 1(Y = \hat{Y}) \):

**KL-projection:**

\[
\min_{q(Y)} \mathcal{D}_{KL}(q(Y) \parallel p_\theta(Y|X)) + \| \mathbb{E}_q[\phi(X, Y)] - b \|_\beta \\
\downarrow
\max_Y \log(p_\theta(Y)) + \log p_N(b|\phi(X, Y)) = -\text{penalty}(\hat{Y})
\]

Use the normal M-step for hard-EM.

**CoDL Objective:**

\[
\max_{\theta, Y} \mathcal{L}(\theta) + \log p_\theta(Y|X) + \log p_N(b|\phi(Y, X))
\]
Types of constraints

$$\arg \max_Y \log p_\theta(Y|X) - \| \phi(X, Y) - b \|_\beta$$

$$\| \phi(X, Y) - b \|_\beta$$
Types of constraints

**Constraint Driven Learning:** Penalized Viterbi

\[
\arg \max_Y \log p_\theta(Y|X) - \| \phi(X, Y) - b \|_\beta
\]

Easy if \( \| \phi(X, Y) - b \|_\beta \) decompose as the model.
Types of constraints

**Constraint Driven Learning:** Penalized Viterbi

\[
\arg\max_Y \log p_{\theta}(Y|X) - \|\phi(X, Y) - b\|_\beta
\]

Easy if \(\|\phi(X, Y) - b\|_\beta\) decompose as the model.

\[
p(Y|X) = \prod_c \psi_c(y_c|X) \quad \text{and} \quad \|\phi(X, Y) - b\|_\beta = \sum_c \delta_c(X, y_c)
\]

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
Types of constraints

**Constraint Driven Learning:** Penalized Viterbi

\[
\arg \max_Y \log p_\theta(Y|X) - \| \phi(X, Y) - b \|_\beta
\]

Easy if \( \| \phi(X, Y) - b \|_\beta \) decompose as the model.

\[
p(Y|X) = \prod_c \psi_c(y_c|X) \quad \text{and} \quad \| \phi(X, Y) - b \|_\beta = \sum_c \delta_c(X, y_c)
\]

Otherwise:

- Beam search
- Integer linear program
- Dual decomposition

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Visual Summary

\[
\log \mathbb{E}[p_N(b|\phi)] \approx \log p_N(b|\mathbb{E}[\phi])
\]
Visual Summary

\[
\log \mathbb{E}[p_N(b|\phi)] \approx \log p_N(b|\mathbb{E}[\phi])
\]
Visual Summary

\[ \log \mathbb{E}[p_N(b|\phi)] \approx \log p_N(b|\mathbb{E}[\phi]) \]

- Generalized Expectation
- Variational approximation; Jensen’s inequality
- Posterior Regularization
Visual Summary

\[ \log E[p_N(b|\phi)] \approx \log p_N(b|E[\phi]) \]

Generalized Expectation \rightarrow variational approximation \rightarrow Posterior Regularization

MAP approximation \rightarrow Constraint Driven Learning

variational approximation; Jensen’s inequality

MAP approximation
Visual Summary

Measurements

Generalized Expectation

Variational approximation; Jensen’s inequality

MAP approximation

Distribution Matching

Quadrianto et al. (2009)

MAP approximation

Constraint Driven Learning

Quadrianto et al. (2009)

MAP approximation

Posterior Regularization

Coupled Semi-Supervised Learning

Carlson et al. (2010)

\[
\log \mathbb{E}[p_N(b|\phi)] \approx \log p_N(b|\mathbb{E}[\phi])
\]
Visual Example: Maximum Likelihood

Model: \[ p(Y|X) = \prod \frac{\exp(y_i x_i \cdot \theta)}{Z(x_i)} \]

Objective: \[ \max_{\theta} \log p_{\theta}(Y_L|X_L) - 0.1\|\theta\|_2^2 \]
Visual Example: Constraint Driven Learning

\[
\max_{\theta, \hat{Y}} \log p_{\theta}(Y_L | X_L) - 0.1\|\theta\|^2_2 \quad \text{s.t.} \quad \phi(\hat{Y}) = 2
\]

where \(\hat{Y}\) are “imagined” labels and \(\phi[\hat{Y}] = \text{count}(+, \hat{Y})\)
Visual Example: Posterior Regularization

\[
\max_{\theta} \log p_{\theta}(Y_L|X_L) - 0.1\|\theta\|^2_2 - D_{\text{KL}}(Q||p_{\theta})
\]

where: \( D_{\text{KL}}(Q||p_{\theta}) = \min_{q} D_{\text{KL}}(q||p_{\theta}) \) s.t. \( E_q[\phi] = 2 \)
A visual comparison of the frameworks

**Objective:** Generalized Expectation Constraints

$$\max_{\theta} \log p_{\theta}(Y_L|X_L) - 0.1\|\theta\|_2^2 - 500\|E_{p_{\theta}}[\phi] - 2\|_2^2$$
Applications Overview

- **Unstructured problems:**
  - Document Classification

- **Sequence problems:**
  - Information Extraction
  - Word Alignment
  - Pos-Induction

- **Tree problems:**
  - Grammar Induction

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Document Classification
but the majority of the film is a convoluted and confusing mess. Characters keep popping up with no explanation, demanding money for deals that occur off-screen.
but the majority of the film is a convoluted and confusing mess. Characters keep popping up with no explanation, demanding money for deals that occur off-screen.
but the majority of the film is a convoluted and confusing mess. Characters keep popping up with no explanation, demanding money for deals that occur off-screen.

but what in the end makes "Toy Story 2" a memorable experience is not the jokes, its multiple parodies or marvelous animation. It is its heart and emotions.
but the majority of the film is a convoluted and confusing mess. Characters keep popping up with no explanation, demanding money for deals that occur off-screen.

but what in the end makes "toy story 2" a memorable experience is not the jokes, its multiple parodies or marvelous animation. It is its heart and emotions.
but the majority of the film is a convoluted and confusing mess. Characters keep popping up with no explanation, demanding money for deals that occur off-screen.

but what in the end makes "toy story 2" a memorable experience is not the jokes, its multiple parodies or marvelous animation. It is its heart and emotions.

**Model:** Max. Entropy Classifier (Logistic Regression)

\[
p_\theta(y|x) = \frac{\exp(\theta \cdot f(x, y))}{\sum_y \exp(\theta \cdot f(x, y))}
\]
but the majority of the film is a convoluted and confusing mess. characters keep popping up with no explanation, demanding money for deals that occur off-screen.

but what in the end makes "toy story 2" a memorable experience is not the jokes, its multiple parodies or marvelous animation. it is its heart and emotions.

**Model:** Max. Entropy Classifier (Logistic Regression)

$$ p_\theta(y|x) = \frac{\exp(\theta \cdot f(x, y))}{\sum_y \exp(\theta \cdot f(x, y))} $$

**One feature for each word / label pair**

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
but the majority of the film is a convoluted and confusing mess. Characters keep popping up with no explanation, demanding money for deals that occur off-screen.

but what in the end makes "toy story 2" a memorable experience is not the jokes, its multiple parodies or marvelous animation. It is its heart and emotions.

Model: Max. Entropy Classifier

\[ p_\theta(y|x) = \frac{\exp(\theta \cdot f(x, y))}{\sum_y \exp(\theta \cdot f(x, y))} \]
but the majority of the film is a convoluted and confusing mess. Characters keep popping up with no explanation, demanding money for deals that occur off-screen.

but what in the end makes "Toy Story 2" a memorable experience is not the jokes, its multiple parodies or marvelous animation. It is its heart and emotions.

**Model:** Max. Entropy Classifier

\[ p_\theta(y|x) = \frac{\exp(\theta \cdot f(x, y))}{\sum_y \exp(\theta \cdot f(x, y))} \]

**Objective:**

\[ \mathcal{L} = \sum_{(x, y) \in D} \log(p_\theta(y|x)) \]
but the majority of the film is a convoluted and confusing mess. Characters keep popping up with no explanation, demanding money for deals that occur off-screen.

but what in the end makes "toy story 2" a memorable experience is not the jokes, its multiple parodies or marvelous animation. It is its heart and emotions.

**Model:** Max. Entropy Classifier

\[ p_\theta(y|x) = \frac{\exp(\theta \cdot f(x, y))}{\sum_y \exp(\theta \cdot f(x, y))} \]

**Objective:**

\[ \mathcal{L} = \sum_{(x,y) \in D} \log(p_\theta(y|x)) \]

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
Document Classification
What if we have no data?
Document Classification
What if we have no data?

but the majority of the film is a convoluted and confusing mess. Characters keep popping up with no explanation, demanding money for deals that occur off-screen.
Document Classification
What if we have no data?

but the majority of the film is a convoluted and confusing mess. Characters keep popping up with no explanation, demanding money for deals that occur off-screen.

but what in the end makes "toy story 2" a memorable experience is not the jokes, its multiple parodies or marvelous animation. It is its heart and emotions.
Document Classification
What if we have no data?

but the majority of the film is a convoluted and confusing mess. Characters keep popping up with no explanation, demanding money for deals that occur off-screen.

but what in the end makes "toy story 2" a memorable experience is not the jokes, its multiple parodies or marvelous animation. It is its heart and emotions.

Objective: $\mathcal{L} = \sum_{(x) \in D} \log(\sum_y p_{\theta}(y|x)) = 0$
Document Classification
What if we have no data?

Objective: \( L = \sum_{(x) \in D} \log(\sum_y p_\theta(y|x)) = 0 \)

Cannot use standard unsupervised learning with ME

but what in the end makes "toy story 2" a memorable experience is not the jokes, its multiple parodies or marvelous animation. It is its heart and emotions.

but the majority of the film is a convoluted and confusing mess. Characters keep popping up with no explanation, demanding money for deals that occur off-screen.
Document Classification
What if we have no data?

but the majority of the film is a convoluted and confusing mess. Characters keep popping up with no explanation, demanding money for deals that occur off-screen.

but what in the end makes "toy story 2" a memorable experience is not the jokes, its multiple parodies or marvelous animation. It is its heart and emotions.

Objective: $\mathcal{L} = \sum_{(x) \in D} \log(\sum_{y} p_{\theta}(y|x)) = 0$

Cannot use standard unsupervised learning with ME
We still have some prior knowledge about the problem.

updated slides: http://sideinfo.wikki.com
Document Classification
What if we have no data?

Objective: $\mathcal{L} = \sum_{(x) \in D} \log(\sum_y p_\theta(y|x)) = 0$

Cannot use standard unsupervised learning with ME
We still have some prior knowledge about the problem

Positive: memorable, marvelous

but the majority of the film is a convoluted and confusing mess. Characters keep popping up with no explanation, demanding money for deals that occur off-screen.

but what in the end makes "toy story 2" a memorable experience is not the jokes, its multiple parodies or marvelous animation. It is its heart and emotions.
Document Classification
What if we have no data?

but the majority of the film is a convoluted and confusing mess. Characters keep popping up with no explanation, demanding money for deals that occur off-screen.

but what in the end makes "toy story 2" a memorable experience is not the jokes, its multiple parodies or marvelous animation. It is its heart and emotions.

Objective: \( \mathcal{L} = \sum_{(x) \in D} \log(\sum_y p_\theta(y|x)) = 0 \)

Cannot use standard unsupervised learning with ME

We still have some prior knowledge about the problem

Positive: memorable, marvelous

Negative: mess

updated slides: http://sideinfo.wikii.com
Document Classification
Labeled features
[Mann & McCallum 07], [Druck et al. 08]
Document Classification
Labeled features
[Mann & McCallum 07], [Druck et al. 08]

• **feature:**

\[
\phi_{w\ell}(x, y) = \begin{cases} 
1 & w \in x \text{ and } y = \ell \\
0 & \text{otherwise}
\end{cases}
\]
Document Classification
Labeled features
[Mann & McCallum 07], [Druck et al. 08]

- **feature:**

\[
\phi_{\omega \ell}(x, y) = \begin{cases} 
1 & \text{if } w \in x \text{ and } y = \ell \\
0 & \text{otherwise}
\end{cases}
\]

- **expectation:** label distribution for docs that contain \( w \)
Document Classification
Labeled features

[Mann & McCallum 07], [Druck et al. 08]

- **feature:**
  \[ \phi_{w \ell}(x, y) \begin{cases} 
    1 & w \in x \text{ and } y = \ell \\
    0 & \text{otherwise} 
  \end{cases} \]

- **expectation:** label distribution for docs that contain \( w \)
  \[ \frac{1}{c_w} \sum_x E_{p_\theta(y|x)} [\phi_w(x, y)] \]
Document Classification
Labeled features
[Mann & McCallum 07], [Druck et al. 08]

• **feature:**
  \[ \phi_{w \ell}(x, y) \begin{cases} 
  1 & w \in x \text{ & } y = \ell \\
  0 & \text{otherwise} 
  \end{cases} \]

• **expectation:** label distribution for docs that contain \( w \)
  \[ \frac{1}{c_w} \sum_{x} E_{p_\theta}(y|x) \left[ \phi_{w}(x, y) \right] \]

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Document Classification
Labeled features
[Mann & McCallum 07], [Druck et al. 08]

• feature:
  \[ \phi_{w \ell}(x, y) = \begin{cases} 1 & w \in x \land y = \ell \\ 0 & \text{otherwise} \end{cases} \]

• expectation: label distribution for docs that contain \( w \)
  \[ \frac{1}{c_w} \sum_x E_{p_{\theta}(y|x)}[\phi_w(x, y)] \]

• GE penalty: KL divergence from target distribution
  \[ D_{KL}(b \mid \mid \frac{1}{c_w} \sum_x E_{p_{\theta}(y|x)}[\phi_w(x, y)]) \]

updated slides: http://sideinfo.wikki.com
User Experiments with Labeled Features
[Druck et al. 08]
User Experiments with Labeled Features

[Druck et al. 08]

PC vs. Mac

- testing accuracy
- labeling time in seconds

updated slides: http://sideinfo.wikii.com
User Experiments with Labeled Features

[Druck et al. 08]

~15 minutes, 100 documents labeled (or skipped): 78% accuracy

updated slides: http://sideinfo.wikii.com
User Experiments with Labeled Features

[Druck et al. 08]

~2 minutes, 100 features labeled (or skipped): 82% accuracy
~15 minutes, 100 documents labeled (or skipped): 78% accuracy

PC vs. Mac

updated slides: http://sideinfo.wikkii.com
User Experiments with Labeled Features

[Druck et al. 08]

~2 minutes, 100 features labeled (or skipped):
82% accuracy

~15 minutes, 100 documents labeled (or skipped):
78% accuracy

targets set with simple heuristic: majority label gets 90% of mass
User Experiments with Labeled Features
[Druck et al. 08]

PC vs. Mac

~2 minutes, 100 features labeled
(82% accuracy)

~15 minutes, 100 documents labeled
(78% accuracy)

targets set with simple heuristic:
majority label gets 90% of mass

complete set of labeled features

<table>
<thead>
<tr>
<th>PC</th>
<th>Mac</th>
</tr>
</thead>
<tbody>
<tr>
<td>dos</td>
<td>mac</td>
</tr>
<tr>
<td>ibm</td>
<td>apple</td>
</tr>
<tr>
<td>hp</td>
<td>quadra</td>
</tr>
<tr>
<td>dx</td>
<td></td>
</tr>
</tbody>
</table>

updated slides: http://sideinfo.wikii.com
Experiments with Labeled Features

[Druck et al. 08]
Experiments with Labeled Features

[Druck et al. 08]

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentiment</td>
<td>50 feat</td>
</tr>
<tr>
<td>webkb</td>
<td>100 feat</td>
</tr>
<tr>
<td>newsgroups</td>
<td>500 feat</td>
</tr>
</tbody>
</table>

Accuracy

- GE (model also contains unlabeled features)
- GE (without unlabeled features)

updated slides: http://sideinfo.wikii.com
Experiments with Labeled Features

[Druck et al. 08]

estimated speed-up over labeling documents

Accuracy

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GE (model also contains unlabeled features)</th>
<th>GE (without unlabeled features)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentiment (50 feat)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>webkb (100 feat)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>newsgroups (500 feat)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Experiments with Labeled Features

[Druck et al. 08]

estimated speed-up over labeling documents

Accuracy

sentiment (50 feat)    webkb (100 feat)    newsgroups (500 feat)

GE (model also contains unlabeled features)    GE (without unlabeled features)

updated slides: http://sideinfo.wikii.com
Experiments with Labeled Features

[Druck et al. 08]

estimated speed-up over labeling documents

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GE (model also contains unlabeled features)</th>
<th>GE (without unlabeled features)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentiment (50 feat)</td>
<td>15x</td>
<td></td>
</tr>
<tr>
<td>webkb (100 feat)</td>
<td>3.5x</td>
<td></td>
</tr>
<tr>
<td>newsgroups (500 feat)</td>
<td>6.5x</td>
<td></td>
</tr>
</tbody>
</table>

Accuracy

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Experiments with Labeled Features

[Druck et al. 08]

- sentiment (50 feat)
- webkb (100 feat)
- newsgroups (500 feat)

GE (model also contains unlabeled features)
GE (without unlabeled features)

estimated speed-up over labeling documents
learning about “unlabeled features” through covariance improves generalization

Accuracy

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GE (model also contains unlabeled features)</th>
<th>GE (without unlabeled features)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentiment</td>
<td>60</td>
<td>60</td>
</tr>
<tr>
<td>webkb</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>newsgroups</td>
<td>75</td>
<td>75</td>
</tr>
</tbody>
</table>

Accuracy

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GE (model also contains unlabeled features)</th>
<th>GE (without unlabeled features)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentiment</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>webkb</td>
<td>75</td>
<td>75</td>
</tr>
<tr>
<td>newsgroups</td>
<td>75</td>
<td>75</td>
</tr>
</tbody>
</table>

estimated speed-up over labeling documents
learning about “unlabeled features” through covariance improves generalization

Accuracy

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GE (model also contains unlabeled features)</th>
<th>GE (without unlabeled features)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentiment</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>webkb</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>newsgroups</td>
<td>80</td>
<td>80</td>
</tr>
</tbody>
</table>

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
Experiments with Labeled Features

[Druck et al. 08]

estimated speed-up over labeling documents

learning about “unlabeled features” through covariance improves generalization

Accuracy

sentiment (50 feat)  webkb (100 feat)  newsgroups (500 feat)

GE (model also contains unlabeled features)

GE (without unlabeled features)

updated slides: http://sideinfo.wikkii.com
Experiments with Labeled Features
[Druck et al. 08]

estimated speed-up over labeling documents

learning about “unlabeled features” through covariance improves generalization

Accuracy

<table>
<thead>
<tr>
<th></th>
<th>GE (model also contains unlabeled features)</th>
<th>GE (without unlabeled features)</th>
</tr>
</thead>
<tbody>
<tr>
<td>sentiment (50 feat)</td>
<td>15x</td>
<td></td>
</tr>
<tr>
<td>webkb (100 feat)</td>
<td>3.5x</td>
<td></td>
</tr>
<tr>
<td>newsgroups (500 feat)</td>
<td>6.5x</td>
<td></td>
</tr>
</tbody>
</table>

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Information Extraction: Example Tasks
Information Extraction: Example Tasks

- **citation extraction:**

Information Extraction: Example Tasks

- **citation extraction:**
  

- **apartment listing extraction:**
  
Information Extraction: Markov Models

- models for **sequence labeling** based IE

- **Hidden Markov Model (HMM):**

\[
p_\theta(y, x) = p_\theta(y_0) \prod_{i=1}^{N} p_\theta(y_i|y_{i-1}) p_\theta(x_i|y_i)
\]

- **Conditional Random Field (CRF):**

\[
p_\theta(y|x) = \frac{1}{Z(x)} \exp(\sum_{i=1}^{N} \theta \cdot f(x, y_{i-1}, y_i))
\]
Information Extraction: Labeled Features

[Mann & McCallum 08], [Liang et al. 09]

apartments example labeled features:

<table>
<thead>
<tr>
<th>ROOMMATES</th>
<th>respectful</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTACT</td>
<td><em>phone</em></td>
</tr>
<tr>
<td>FEATURES</td>
<td>laundry</td>
</tr>
</tbody>
</table>
Information Extraction: Labeled Features

[Mann & McCallum 08], [Liang et al. 09]

apartments example labeled features:

| ROOMMATES | respectful |
| CONTACT   | *phone* |
| FEATURES  | laundry |

• feature: $\phi_q(x, y_i, i)$
Information Extraction: Labeled Features

[Mann & McCallum 08], [Liang et al. 09]

| ROOMMATES | respectful |
| CONTACT | *phone* |
| FEATURES | laundry |

apartments example labeled features:

- **feature:** $\phi_q(x, y_i, i)$

- **expectation:**
  $$\frac{1}{c_q} \sum_x \sum_i \mathbb{E}_{p(\theta)(y_i|x)} [\phi_q(x, y_i, i)]$$
Information Extraction: Labeled Features
[Haghighi & Klein 06], [Mann & McCallum 08], [Liang et al. 09]

apartment listing extraction

Prototype
GE (KL)
Measurements/PR

Accuracy

updated slides: http://sideinfo.wikkii.com
Information Extraction: Labeled Features
[Haghighi & Klein 06], [Mann & McCallum 08], [Liang et al. 09]

apartment listing extraction

Prototype
GE (KL)
Measurements/PR

Accuracy

85
80
75
70
65

0 labeled ex 10 labeled ex 100 labeled ex

updated slides: http://sideinfo.wikkii.com
Information Extraction: Labeled Features
[Haghighi & Klein 06], [Mann & McCallum 08], [Liang et al. 09]

apartment listing extraction

Accuracy

<table>
<thead>
<tr>
<th>0 labeled ex</th>
<th>10 labeled ex</th>
<th>100 labeled ex</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prototype</td>
<td>GE (KL)</td>
<td>Measurements/PR</td>
</tr>
</tbody>
</table>

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Information Extraction: Labeled Features
[Haghighi & Klein 06], [Mann & McCallum 08], [Liang et al. 09]

apartment listing extraction

- accurate with constraints alone

Accuracy

- Prototype
- GE (KL)
- Measurements/PR

updated slides: http://sideinfo.wikii.com
Information Extraction: Labeled Features
[Haghighi & Klein 06], [Mann & McCallum 08], [Liang et al. 09]

apartment listing extraction

- accurate with constraints alone

Accuracy

Prototype | GE (KL) | Measurements/PR

<table>
<thead>
<tr>
<th>0 labeled ex</th>
<th>10 labeled ex</th>
<th>100 labeled ex</th>
</tr>
</thead>
<tbody>
<tr>
<td>65</td>
<td>70</td>
<td>75</td>
</tr>
<tr>
<td>70</td>
<td>98</td>
<td>85</td>
</tr>
<tr>
<td>75</td>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>80</td>
<td>85</td>
<td>85</td>
</tr>
</tbody>
</table>

updated slides: http://sideinfo.wikkii.com
Information Extraction: Labeled Features
[Haghighi & Klein 06], [Mann & McCallum 08], [Liang et al. 09]

apartment listing extraction

- accurate with constraints alone
- outperform fully supervised with constraints and labeled data

Prototype
GE (KL)
Measurements/PR

Accuracy

supervised CRF (100) [MM08]

updated slides: http://sideinfo.wikki.com
Limitations of Markov Models
Limitations of Markov Models

- **predicted:**

Limitations of Markov Models


- prediction has two **author** and two **title** segments:
Limitations of Markov Models


- prediction has two **author** and two **title** segments:

- **error #1:** Neuhold, Ed. should be **editor**

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Limitations of Markov Models

• predicted:


• prediction has two author and two title segments:

• error #1: Neuhold, Ed. should be editor

• error #2: North-Holland Pub. Co., should be publisher

updated slides: http://sideinfo.wikii.com
Limitations of Markov Models

- **predicted:**

  

- prediction has two **author** and two **title** segments:
  - **error #1:** Neuhold, Ed. should be **editor**
  - **error #2:** North-Holland Pub. Co., should be **publisher**

  - A Markov model cannot represent that at most one segment of each type appears in each reference.
Long-Range Constraints
[Chang et al. 07] [Bellare et al. 09]

updated slides: http://sideinfo.wikii.com
Long-Range Constraints
[Chang et al. 07] [Bellare et al. 09]

• **feature:** “Each field is a contiguous sequence of tokens and appears at most once in a citation.”
Long-Range Constraints
[Chang et al. 07] [Bellare et al. 09]

- **feature:** “Each field is a contiguous sequence of tokens and appears at most once in a citation.”
- Does not decompose (beam search)
Long-Range Constraints
[Chang et al. 07] [Bellare et al. 09]

• **feature:** “Each field is a contiguous sequence of tokens and appears at most once in a citation.”

• Does not decompose (beam search)

• **Constrain:** $E_q[\phi(x, y)] \leq 1$
Long-Range Constraints
[Chang et al. 07] [Bellare et al. 09]

- **feature:** “Each field is a contiguous sequence of tokens and appears at most once in a citation.”
- Does not decompose (beam search)

- **Constrain:** $E_q[\phi(x, y)] \leq 1$

- **additional constraints:** 10 labeled features such as:
  - pages $\rightarrow$ pages
  - proc. $\rightarrow$ booktitle

updated slides: [http://sideinfo.wikiki.com](http://sideinfo.wikiki.com)
Long-Range Constraints
[Chang et al. 07] [Bellare et al. 09]

- CRF
- HMM
- CRF + PR
- HMM + CODL

Accuracy

updated slides: http://sideinfo.wikii.com
Long-Range Constraints
[Chang et al. 07] [Bellare et al. 09]

- CRF
- CRF + PR
- HMM
- HMM + CODL

Accuracy

 updated slides:  http://sideinfo.wikkii.com  101
Long-Range Constraints
[Chang et al. 07] [Bellare et al. 09]

- CRF
- CRF + PR
- HMM
- HMM + CODL

Accuracy

5 labeled examples

20 labeled examples

updated slides: http://sideinfo.wikii.com
Long-Range Constraints
[Chang et al. 07] [Bellare et al. 09]

Accuracy

5 labeled examples

20 labeled examples

CRF    CRF + PR
HMM    HMM + CODL

updated slides: http://sideinfo.wikii.com
Long-Range Constraints
[Chang et al. 07] [Bellare et al. 09]

Accuracy

CRF  CRF + PR  HMM  HMM + CODL

5 labeled examples

20 labeled examples

updated slides: http://sideinfo.wikkii.com
Long-Range Constraints
[Chang et al. 07] [Bellare et al. 09]

Accuracy

- CRF
- CRF + PR
- HMM
- HMM + CODL

Updated slides: http://sideinfo.wikkii.com
Long-Range Constraints
[Chang et al. 07] [Bellare et al. 09]

Constraints improve both CRF (PR) and HMM (CODL)

Accuracy

CRF  CRF + PR  HMM  HMM + CODL

5 labeled examples  20 labeled examples

updated slides: http://sideinfo.wikii.com
Word Alignments
Unsupervised
A career with the European institutions must become more attractive.

Uma carreira nas instituições europeias tem de se tornar mais atractiva.
Word Alignments
Unsupervised

A career with the European institutions must become more attractive.

Uma carreira nas instituições europeias tem de se tornar mais atractiva.

updated slides: http://sideinfo.wikki.com
Word Alignments
HMM Model
Word Alignments

HMM Model

\[ p_\theta(x_t|y_t) \]

\[ p_\theta(y_t|y_{t-1}) \]

we
know
the
way

1
sabemos
el
camino
null

2

3

0
Word Alignments
HMM Model

$P_{\theta}(y_t|y_{t-1})$ : Distortion Probabilities
$P_{\theta}(x_t|y_t)$ : Translation Probabilities
Word Alignments
HMM Model

\[ p_\theta(y_t|y_{t-1}) \]  
\[ p_\theta(x_t|y_t) \]

\[ p_\theta(x_t|y_t) \]: Translation Probabilities
\[ p_\theta(y_t|y_{t-1}) \]: Distortion Probabilities

Alignments are directional 1-n

updated slides: http://sideinfo.wikii.com

updated slides: http://sideinfo.wikii.com
Word Alignments

\[ p_\theta(y|x) = 0.7 \]
Word Alignments

\[ p_\theta(y|x) = 0.7 \]

\[ p_\theta(y|x) = 0.2 \]

updated slides: http://sideinfo.wikii.com
Word Alignments

$p_\theta(y|x) = 0.7$

$p_\theta(y|x) = 0.2$

$p_\theta(y|x) = 0.1$
Word Alignments

\[ p_\theta(y|x) = 0.7 \]

\[ p_\theta(y|x) = 0.2 \]

\[ p_\theta(y|x) = 0.1 \]

All other alignments are have zero probability......

updated slides: http://sideinfo.wikkii.com
Word Alignments

\[ \begin{align*}
  p_\theta(y|x) &= 0.7 \\
  p_\theta(y|x) &= 0.2 \\
  p_\theta(y|x) &= 0.1
\end{align*} \]

All other alignments are have zero probability......

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Word Alignments

\[ p_\theta(y|x) = 0.7 \]

\[ p_\theta(y|x) = 0.2 \]

\[ p_\theta(y|x) = 0.1 \]

All other alignments are have zero probability......

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Word Alignments

\[ p_\theta(y|x) = 0.7 \]

\[ p_\theta(y|x) = 0.2 \]

\[ p_\theta(y|x) = 0.1 \]

All other alignments are have zero probability......
All other alignments are have zero probability......
Word Alignments

\[ p_\theta(y|x) = 0.7 \]

\[ p_\theta(y|x) = 0.2 \]

\[ p_\theta(y|x) = 0.1 \]

All other alignments are have zero probability......

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0.1</td>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>S2</td>
<td>0.9</td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>S3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Word Alignments

\[ p_\theta(y|x) = 0.7 \]

\[ p_\theta(y|x) = 0.2 \]

\[ p_\theta(y|x) = 0.1 \]

All other alignments are have zero probability......

<table>
<thead>
<tr>
<th></th>
<th>T1</th>
<th>T2</th>
<th>T3</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>0.1</td>
<td>0.7</td>
<td>0</td>
</tr>
<tr>
<td>S2</td>
<td>0.9</td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>S3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

updated slides: http://sideinfo.wikki.com
Word Alignments

What is wrong with this model
Word Alignments
What is wrong with this model

Alignments are directional 1-n
Word Alignments
What is wrong with this model

Alignments are directional 1-n

jugaban de una manera animada y muy cordial

it was an animated, very convivial game
Word Alignments
What is wrong with this model

Alignments are directional 1-n

jugaban de una manera animada y muy cordial

it was an animated, very convivial game

Garbage Collector Effect

updated slides: http://sideinfo.wikkii.com
Word Alignments
What is wrong with this model

Alignments are directional 1-n

jugaban de una manera animada y muy cordial

it was an animated, very convivial game

Garbage Collector Effect
Word Alignments

[Graça et al. 10]
Word Alignments

[Graça et al. 10]

- **Bijectivity constraints:**
  - Each word should align to at most one other word
Word Alignments

[Graça et al. 10]

- **Bijectivity constraints:**
  - Each word should align to at most one other word
Word Alignments

[Graça et al. 10]

- **Bijectivity constraints:**
  - Each word should align to at most one other word

- **Symmetry constraints:**
  - Directional models should agree
Bijectivity Constraints

[Graça et al. 10]
Bijectivity Constraints

[Graça et al. 10]

Updated slides: http://sideinfo.wikki.com
Bijectivity Constraints

[Graca et al. 10]

Feature: $\phi(x, y) = \sum_{i=1}^{N} 1(y_i = m)$
Bijectivity Constraints

[Graca et al. 10]

Feature: \( \phi(x, y) = \sum_{i=1}^{N} 1(y_i = m) \)
Bijectivity Constraints

[Grăca et al. 10]

**Feature:** \( \phi(x, y) = \sum_{i=1}^{N} 1(y_i = m) \)
Bijectivity Constraints

[Graca et al. 10]

**Feature:** \( \phi(x, y) = \sum_{i=1}^{N} 1(y_i = m) \)
**Bijectivity Constraints**

[Graca et al. 10]

**Feature:** \( \phi(x, y) = \sum_{i=1}^{N} 1(y_i = m) \)

**Constraint:** \( E_q[\phi(x, y)] \leq 1 \)

updated slides: http://sideinfo.wikkii.com
Bijectivity Constraints

[Graca et al. 10]

Feature: \( \phi(x, y) = \sum_{i=1}^{N} 1(y_i = m) \)

Constraint: \( E_q[\phi(x, y)] \leq 1 \)

updated slides: http://sideinfo.wikki.com
Symmetry Constraints
[Graça et al. 10]

updated slides: http://sideinfo.wikki.com
Symmetry Constraints

[Graca et al. 10]

Forward: $\overrightarrow{p}_\theta(y|x)$
Symmetry Constraints
[Graça et al. 10]

**Forward:** $\bar{p}_\theta(y|x)$

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>no</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1</td>
<td>hay</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>estadísticas</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Symmetry Constraints
[Graça et al. 10]

Forward: $\overrightarrow{p}_\theta(y|x)$

Backward: $\overleftarrow{p}_\theta(y|x)$

updated slides: http://sideinfo.wikkii.com
Symmetry Constraints

[Graca et al. 10]

**Forward:** $\mathbf{P}_\theta (y | x)$

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

no statistical data exists.

**Backward:** $\mathbf{P}_\theta (y | x)$

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

no statistical data exists.

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Symmetry Constraints

[Graça et al. 10]

**Forward:** \( \overrightarrow{p}_\theta(y|x) \)

\[ 
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
0 & \cdot & \cdot & \cdot & \cdot & \text{no} \\
1 & \cdot & \cdot & \cdot & \cdot & \text{hay} \\
2 & \cdot & \cdot & \cdot & \cdot & \text{estadísticas} \\
3 & \cdot & \cdot & \cdot & \cdot & \cdot \\
\end{array} 
\]

**Backward:** \( \overleftarrow{p}_\theta(y|x) \)

\[ 
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 \\
\cdot & \cdot & \cdot & \cdot & \cdot \\
0 & \cdot & \cdot & \cdot & \cdot & \text{no} \\
1 & \cdot & \cdot & \cdot & \cdot & \text{hay} \\
2 & \cdot & \cdot & \cdot & \cdot & \text{estadísticas} \\
3 & \cdot & \cdot & \cdot & \cdot & \cdot \\
\end{array} 
\]

**Joint Model:** \( p_\theta(y|x) = \frac{1}{2} \overrightarrow{p}_\theta(y|x) + \frac{1}{2} \overleftarrow{p}_\theta(y|x) \)
Symmetry Constraints

[Graca et al. 10]

**Forward:** \( \overrightarrow{p_\theta(y|x)} \)

**Backward:** \( \overleftarrow{p_\theta(y|x)} \)

**Joint Model:**

\[
p_\theta(y|x) = \frac{1}{2} \overrightarrow{p_\theta(y|x)} + \frac{1}{2} \overleftarrow{p_\theta(y|x)}
\]

\[
y = \overrightarrow{y} \cup \overleftarrow{y}
\]

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
Symmetry Constraints
[Graça et al. 10]

**Forward:** \( \overrightarrow{p_\theta(y|x)} \)

**Backward:** \( \overleftarrow{p_\theta(y|x)} \)

**Joint Model:**
\[
p_\theta(y|x) = \frac{1}{2} \overrightarrow{p_\theta(y|x)} + \frac{1}{2} \overleftarrow{p_\theta(y|x)} = \begin{cases} 
\overrightarrow{p_\theta(y|x)} & y \in \overrightarrow{y} \\
0 & y \in \overleftarrow{y}
\end{cases}
\]

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
Symmetry Constraints

[Graça et al. 10]
Symmetry Constraints

[Graca et al. 10]

Joint Model: \[ p_\theta(y|x) = \frac{1}{2} \overrightarrow{p}_\theta(y|x) + \frac{1}{2} \overleftarrow{p}_\theta(y|x) \]

Forward: \[ \overrightarrow{p}_\theta(y|x) \]

\[
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 \\
0 & \cdot & \cdot & \cdot & \cdot & \text{no}
\end{array}
\]

\[
\begin{array}{cccccc}
1 & \cdot & \cdot & \cdot & \cdot & \text{hay}
\end{array}
\]

\[
\begin{array}{cccccc}
2 & \cdot & \cdot & \cdot & \cdot & \text{estadísticas}
\end{array}
\]

\[
\begin{array}{cccccc}
3 & \cdot & \cdot & \cdot & \cdot & \text{data exists}
\end{array}
\]

Backward: \[ \overleftarrow{p}_\theta(y|x) \]

\[
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 \\
0 & \cdot & \cdot & \cdot & \cdot & \text{no}
\end{array}
\]

\[
\begin{array}{cccccc}
1 & \cdot & \cdot & \cdot & \cdot & \text{hay}
\end{array}
\]

\[
\begin{array}{cccccc}
2 & \cdot & \cdot & \cdot & \cdot & \text{estadísticas}
\end{array}
\]

\[
\begin{array}{cccccc}
3 & \cdot & \cdot & \cdot & \cdot & \text{data exists}
\end{array}
\]

updated slides: http://sideinfo.wikkii.com
Symmetry Constraints

[Graça et al. 10]

Joint Model: \( p_\theta(y|x) = \frac{1}{2} \overrightarrow{p}_\theta(y|x) + \frac{1}{2} \overleftarrow{p}_\theta(y|x) \)

Forward: \( \overrightarrow{p}_\theta(y|x) \)

Backward: \( \overleftarrow{p}_\theta(y|x) \)

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Symmetry Constraints

[Graca et al. 10]

**Joint Model:**

\[ p_\theta(y|x) = \frac{1}{2} p_\theta(y|x) + \frac{1}{2} \overrightarrow{p}_\theta(y|x) \]

**Forward:** \( \overrightarrow{p}_\theta(y|x) \)

**Backward:** \( \overleftarrow{p}_\theta(y|x) \)

**Feature:**

\[ \phi(x, y) = \begin{cases} 
+1 & y \in \overrightarrow{y} \text{ and } \overrightarrow{y}_i = j \\
-1 & y \in \overleftarrow{y} \text{ and } \overleftarrow{y}_j = i \\
0 & \text{otherwise}
\end{cases} \]

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Symmetry Constraints

[Graca et al. 10]

Joint Model: \( p_\theta(y|x) = \frac{1}{2} \overrightarrow{p_\theta}(y|x) + \frac{1}{2} \overleftarrow{p_\theta}(y|x) \)

Forward: \( \overrightarrow{p_\theta}(y|x) \)

Backward: \( \overleftarrow{p_\theta}(y|x) \)

Feature: \( \phi(x, y) = \begin{cases} +1 & y \in \overrightarrow{y} \text{ and } \overrightarrow{y}_i = j \\ -1 & y \in \overleftarrow{y} \text{ and } \overleftarrow{y}_j = i \\ 0 & \text{otherwise} \end{cases} \)

(updated slides: http://sideinfo.wikki.com)
Symmetry Constraints

[Graça et al. 10]

Joint Model:
\[ p_\theta(y|x) = \frac{1}{2} \overrightarrow{p}_\theta(y|x) + \frac{1}{2} \overleftarrow{p}_\theta(y|x) \]

Forward:
\[ \overrightarrow{p}_\theta(y|x) \]

\[
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 & \\
0 & \cdot & \cdot & \cdot & \cdot & \cdot \\
1 & \cdot & \cdot & \cdot & \cdot & \cdot \\
2 & \cdot & \cdot & \cdot & \cdot & \cdot \\
3 & \cdot & \cdot & \cdot & \cdot & \cdot \\
\end{array}
\]

Backward:
\[ \overleftarrow{p}_\theta(y|x) \]

\[
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 & \\
0 & \cdot & \cdot & \cdot & \cdot & \cdot \\
1 & \cdot & \cdot & \cdot & \cdot & \cdot \\
2 & \cdot & \cdot & \cdot & \cdot & \cdot \\
3 & \cdot & \cdot & \cdot & \cdot & \cdot \\
\end{array}
\]

Feature:
\[ \phi(x, y) = \begin{cases} 
+1 & y \in \overrightarrow{y} \text{ and } \overrightarrow{y}_i = j \\
-1 & y \in \overleftarrow{y} \text{ and } \overleftarrow{y}_j = i \\
0 & \text{otherwise} 
\end{cases} \]

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Symmetry Constraints

[Graca et al. 10]

Joint Model: \( p_\theta(y|x) = \frac{1}{2} \overrightarrow{p_\theta}(y|x) + \frac{1}{2} \overleftarrow{p_\theta}(y|x) \)

Forward: \( \overrightarrow{p_\theta}(y|x) \)

Backward: \( \overleftarrow{p_\theta}(y|x) \)

Feature: \( \phi(x, y) = \begin{cases} +1 & y \in \overrightarrow{y} \text{ and } \overrightarrow{y}_i = j \\ -1 & y \in \overleftarrow{y} \text{ and } \overleftarrow{y}_j = i \\ 0 & \text{otherwise} \end{cases} \)

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
Symmetry Constraints

[Graca et al. 10]

Joint Model: \( p_\theta(y|x) = \frac{1}{2} \overrightarrow{p}_\theta(y|x) + \frac{1}{2} \overleftarrow{p}_\theta(y|x) \)

Forward: \( \overrightarrow{p}_\theta(y|x) \)

Backward: \( \overleftarrow{p}_\theta(y|x) \)

Feature: \( \phi(x, y) = \begin{cases} +1 & y \in \overrightarrow{y} \text{ and } \overrightarrow{y}_i = j \\ -1 & y \in \overleftarrow{y} \text{ and } \overleftarrow{y}_j = i \\ 0 & \text{otherwise} \end{cases} \)

updated slides: http://sideinfo.wikiki.com
Symmetry Constraints

[Graca et al. 10]

Joint Model: \( p_\theta(y|x) = \frac{1}{2} \overrightarrow{p}_\theta(y|x) + \frac{1}{2} \overleftarrow{p}_\theta(y|x) \)

Forward: \( \overrightarrow{p}_\theta(y|x) \)

Backward: \( \overleftarrow{p}_\theta(y|x) \)

Feature: \( \phi(x, y) = \begin{cases} +1 & y \in \overrightarrow{y} \text{ and } \overrightarrow{y}_i = j \\ -1 & y \in \overleftarrow{y} \text{ and } \overleftarrow{y}_j = i \\ 0 & \text{otherwise} \end{cases} \)

Constraint: \( E_q[\phi(x, y)] = 0 \)

updated slides: http://sideinfo.wikki.com
Symmetry Constraints

[Graça et al. 10]
Symmetry Constraints

[Graca et al. 10]

Before projection:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_s(z)$ = argmin $q(z) \in Q_s$ $KL[q_s(z)</td>
<td></td>
<td>p_\theta(t(z</td>
<td>x))]$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\vec{p}_\theta(y</td>
<td>x)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\overleftarrow{p}_\theta(y</td>
<td>x)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>1</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

no statistical data exists
Symmetry Constraints
[Graca et al. 10]

Before projection:

\[ \overrightarrow{p_{\theta}(y|x)} \]

\[
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 \\
0 & \cdot & \cdot & \cdot & \cdot & \cdot \\
1 & \cdot & \cdot & \cdot & \cdot & \text{no}
\end{array}
\]

\[
\begin{array}{cccccc}
2 & \cdot & \cdot & \cdot & \cdot & \text{hay}
\end{array}
\]

\[
\begin{array}{cccccc}
3 & \cdot & \cdot & \cdot & \cdot & \text{estadísticas}
\end{array}
\]

\[
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 \\
0 & \cdot & \cdot & \cdot & \cdot & \cdot \\
1 & \cdot & \cdot & \cdot & \cdot & \text{no}
\end{array}
\]

\[
\begin{array}{cccccc}
2 & \cdot & \cdot & \cdot & \cdot & \text{hay}
\end{array}
\]

\[
\begin{array}{cccccc}
3 & \cdot & \cdot & \cdot & \cdot & \text{estadísticas}
\end{array}
\]

After projection:

\[ \overrightarrow{q(y)} \]

\[
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 \\
0 & \cdot & \cdot & \cdot & \cdot & \cdot \\
1 & \cdot & \cdot & \cdot & \cdot & \text{no}
\end{array}
\]

\[
\begin{array}{cccccc}
2 & \cdot & \cdot & \cdot & \cdot & \text{hay}
\end{array}
\]

\[
\begin{array}{cccccc}
3 & \cdot & \cdot & \cdot & \cdot & \text{estadísticas}
\end{array}
\]

\[
\begin{array}{cccccc}
0 & 1 & 2 & 3 & 4 \\
0 & \cdot & \cdot & \cdot & \cdot & \cdot \\
1 & \cdot & \cdot & \cdot & \cdot & \text{no}
\end{array}
\]

\[
\begin{array}{cccccc}
2 & \cdot & \cdot & \cdot & \cdot & \text{hay}
\end{array}
\]

\[
\begin{array}{cccccc}
3 & \cdot & \cdot & \cdot & \cdot & \text{estadísticas}
\end{array}
\]

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
Word Alignments
Decoding

updated slides: http://sideinfo.wikii.com
Word Alignments
Decoding

updated slides: http://sideinfo.wikkii.com
Word Alignments
Decoding

Posterior Decoding

\[ p_\theta(y_t|x_t) > \delta \]
Word Alignments
Decoding

Posterior Decoding

\[ p_\theta(y_t | x_t) > \delta \]

Precision/Recall curves

updated slides: http://sideinfo.wikii.com
Results

[Graca et al. 10]
Results

[Graça et al. 10]

• Fix recall according to baseline model
Results

[Graça et al. 10]

• Fix recall according to baseline model
• Measure precision
Results

[Graça et al. 10]

• Fix recall according to baseline model
• Measure precision
Results

[Graça et al. 10]

- Fix recall according to baseline model
- Measure precision

Precision/Recall curves

• Fix recall according to baseline model
• Measure precision
Results

[Graca et al. 10]

<table>
<thead>
<tr>
<th>Languages</th>
<th>HMM</th>
<th>B-HMM</th>
<th>S-HMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>En-Pt</td>
<td>85.0</td>
<td>86.2</td>
<td>85.0</td>
</tr>
<tr>
<td>Pt-En</td>
<td>67.5</td>
<td>74.4</td>
<td></td>
</tr>
<tr>
<td>Pt-Fr</td>
<td>71.3</td>
<td>73.0</td>
<td></td>
</tr>
<tr>
<td>Fr-Pt</td>
<td>77.8</td>
<td>82.4</td>
<td></td>
</tr>
<tr>
<td>En-Es</td>
<td>86.3</td>
<td>87.9</td>
<td></td>
</tr>
<tr>
<td>Es-En</td>
<td>88.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Es-Fr</td>
<td>87.2</td>
<td>84.6</td>
<td>87.2</td>
</tr>
<tr>
<td>Fr-Es</td>
<td>89.1</td>
<td>88.9</td>
<td></td>
</tr>
<tr>
<td>Pt-Es</td>
<td>84.0</td>
<td>86.5</td>
<td></td>
</tr>
<tr>
<td>Es-Pt</td>
<td>82.4</td>
<td>82.5</td>
<td>91.8</td>
</tr>
<tr>
<td>En-Fr</td>
<td>79.8</td>
<td>90.1</td>
<td>93.4</td>
</tr>
<tr>
<td>Fr-En</td>
<td>76.3</td>
<td>90.8</td>
<td>91.6</td>
</tr>
</tbody>
</table>

updated slides: http://sideinfo.wikkii.com
POS Induction

Text
A career with the European institutions must become more attractive. Too many young, new....

updated slides: http://sideinfo.wikkii.com
POS Induction

Text
A career with the European institutions must become more attractive. Too many young, new....

updated slides: http://sideinfo.wikii.com
A career with the European institutions must become more attractive. Too many young, new....

Cluster Words

A career with the European institutions must become more attractive. Too many young, new...
PoS Induction
HMM Model

\[ p_\theta(y_t | y_{t-1}) \]

\[ p_\theta(x_t | y_t) \]

S3 \rightarrow S2 \rightarrow S1 \rightarrow S4

a, run, into, town

updated slides: http://sideinfo.wikii.com
PoS Induction
HMM Model

$\Pr(y_t | y_{t-1})$

$\Pr(x_t | y_t)$

$\Pr(y_t | y_{t-1})$ : Transition Probabilities: Multinomial
PoS Induction
HMM Model

\[ p_\theta(y_t | y_{t-1}) \]

\[ p_\theta(x_t | y_t) \]

\[ p_\theta(y_t | y_{t-1}) : \text{Transition Probabilities: Multinomial} \]

\[ p_\theta(x_t | y_t) : \text{Observation Probabilities: Multinomial} \]
PoS Induction
What is wrong with this model
PoS Induction
What is wrong with this model

avg. degree = 10000
avg. degree = 1.5

- DT - speak
- VB - the
- NN - run
- JJ - offensive
- - romantic

Model: Hidden Markov model
Training: Fully unsupervised
Prior Knowledge: few POS tags per word type
PoS Induction
What is wrong with this model

avg. degree = 10000
avg. degree = 1.5

Distribution of word ambiguity

Supervised HMM

The brown fox jumps over the fence.

Car offensive romantic

avg. degree = 1.5

updated slides: http://sideinfo.wikii.com
PoS Induction
Measuring Tag Ambiguity
[Graça et al. 09]
PoS Induction
Measuring Tag Ambiguity

[Graça et al. 09]

• Pick a particular word type: run
  • Stack all occurrences
PoS Induction
Measuring Tag Ambiguity
[Graça et al. 09]

• Pick a particular word type: run
  • Stack all occurrences

a run into town.
of the mile run.
  run gold.
  run errands.
  run for mayor.
Pick a particular word type: \texttt{run}

Stack all occurrences

Calculate posterior probability tag:
\[
pr(s_t = \textit{noun} | w_t = \texttt{run})
\]
PoS Induction
Measuring Tag Ambiguity
[Graça et al. 09]

• Pick a particular word type: run
  • Stack all occurrences
  • Calculate posterior probability tag:
    \[ pr(s_t = noun | w_t = run) \]

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>V</th>
<th>ADJ</th>
<th>Prep</th>
<th>ADV</th>
</tr>
</thead>
<tbody>
<tr>
<td>a run into town.</td>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>of the mile run.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>run gold.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>run errands.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>run for mayor.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
PoS Induction
Measuring Tag Ambiguity
[Graça et al. 09]

a run into town.
of the mile run.
run gold.
run errands.
run for mayor.

<table>
<thead>
<tr>
<th>N</th>
<th>V</th>
<th>ADJ</th>
<th>Prep</th>
<th>ADV</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

• Pick a particular word type: run
• Stack all occurrences
• Calculate posterior probability tag:
  \[ pr(s_t = \text{noun} | w_t = \text{run}) \]
PoS Induction
Measuring Tag Ambiguity
[Graça et al. 09]

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>V</th>
<th>ADJ</th>
<th>Prep</th>
<th>ADV</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>a run</td>
<td>0.9</td>
<td>0.1</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>into</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>town.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>of the</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
<td>0.0</td>
<td>0.1</td>
<td>1.0</td>
</tr>
<tr>
<td>mile</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>run</td>
<td>0.1</td>
<td>0.3</td>
<td>0.0</td>
<td>0.6</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>gold.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>run</td>
<td>0.3</td>
<td>0.6</td>
<td>0.0</td>
<td>0.0</td>
<td>0.1</td>
<td>1.0</td>
</tr>
<tr>
<td>errands.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>run</td>
<td>0.3</td>
<td>0.7</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
<tr>
<td>for</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mayor.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Pick a particular word type: run
- Stack all occurrences
- Calculate posterior probability tag:
- Stack together all occurrences

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
### PoS Induction

#### Measuring Tag Ambiguity

[Graça et al. 09]

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>V</th>
<th>ADJ</th>
<th>Prep</th>
<th>ADV</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>a run into town.</td>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>of the mile run.</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>run gold.</td>
<td>0.1</td>
<td>0.3</td>
<td>0</td>
<td>0.6</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>run errands.</td>
<td>0.3</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>run for mayor.</td>
<td>0.3</td>
<td>0.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- Pick a particular word type: **run**
- Stack all occurrences
- Calculate posterior probability tag:
- Stack together all occurrences
- Take the maximum for each tag

---

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
PoS Induction
Measuring Tag Ambiguity

[Graca et al. 09]

• Pick a particular word type: run
  • Stack all occurrences
• Calculate posterior probability tag:
  • Stack together all occurrences
  • Take the maximum for each tag
  • Sum the maxes

updated slides: http://sideinfo.wikii.com
PoS Induction
Measuring Tag Ambiguity
[Graça et al. 09]

<table>
<thead>
<tr>
<th>N</th>
<th>V</th>
<th>ADJ</th>
<th>Prep</th>
<th>ADV</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

a run into town.

... of the mile run.

  run gold.

  run errands.

  run for mayor.

updated slides: http://sideinfo.wikkii.com
PoS Induction
Measuring Tag Ambiguity
[Graça et al. 09]

Using the same tag is free

updated slides: http://sideinfo.wikki.com
PoS Induction
Measuring Tag Ambiguity
[Graça et al. 09]

<table>
<thead>
<tr>
<th>N</th>
<th>V</th>
<th>ADJ</th>
<th>Prep</th>
<th>ADV</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>0.1</td>
<td>0.3</td>
<td>0</td>
<td>0.6</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- Using the same tag is free
- Picking a different tag costs

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
PoS Induction
Measuring Tag Ambiguity
[Graça et al. 09]

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>V</th>
<th>ADJ</th>
<th>Prep</th>
<th>ADV</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>a run into town.</td>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>... of the mile run.</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>run gold.</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0.6</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>run errands.</td>
<td>0.3</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>1</td>
</tr>
</tbody>
</table>

• Using the same tag is free
• Picking a different tag costs

updated slides: http://sideinfo.wikki.com
PoS Induction
Measuring Tag Ambiguity
[Graça et al. 09]

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>V</th>
<th>ADJ</th>
<th>Prep</th>
<th>ADV</th>
</tr>
</thead>
<tbody>
<tr>
<td>run</td>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>run</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>run</td>
<td>0.1</td>
<td>0.3</td>
<td>0</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>run</td>
<td>0.3</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>run</td>
<td>0.3</td>
<td>0.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Using the same tag is free
• Picking a different tag costs
• Bound the Sum of the Maxes

Sum

Max

Sum

2.4
### PoS Induction

#### Measuring Tag Ambiguity

[Graça et al. 09]

<table>
<thead>
<tr>
<th>N</th>
<th>V</th>
<th>ADJ</th>
<th>Prep</th>
<th>ADV</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>0.2</td>
<td>0.4</td>
<td>0.1</td>
<td>0.2</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>0.3</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>1</td>
</tr>
<tr>
<td>0.3</td>
<td>0.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- Using the same tag is free
- Picking a different tag costs
- Bound the Sum of the Maxes
- Outliers easier to eliminate

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
PoS Induction
Measuring Tag Ambiguity

[Graça et al. 09]
PoS Induction
Measuring Tag Ambiguity
[Graça et al. 09]

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>V</th>
<th>ADJ</th>
<th>Prep</th>
<th>ADV</th>
</tr>
</thead>
<tbody>
<tr>
<td>a run</td>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>... of</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>run</td>
<td>0.1</td>
<td>0.1</td>
<td>0</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>run</td>
<td>0.3</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>run</td>
<td>0.3</td>
<td>0.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Feature: $\phi_{wti}$ : Word type $w$ has hidden state $t$ at occurrence $i$
PoS Induction
Measuring Tag Ambiguity
[Graça et al. 09]

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>V</th>
<th>ADJ</th>
<th>Prep</th>
<th>ADV</th>
</tr>
</thead>
<tbody>
<tr>
<td>a run into town.</td>
<td>0.9</td>
<td>0.1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>... of the mile run.</td>
<td>0.7</td>
<td>0.1</td>
<td>0.1</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>run gold.</td>
<td>0.1</td>
<td>0.3</td>
<td>0</td>
<td>0.6</td>
<td>0</td>
</tr>
<tr>
<td>run errands.</td>
<td>0.3</td>
<td>0.6</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
</tr>
<tr>
<td>run for mayor.</td>
<td>0.3</td>
<td>0.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ \text{Max} \]
\[ \text{Sum} \]

\[ \begin{array}{lllll}
0.9 & 0.7 & 0.1 & 0.6 & 0.2 \\
\end{array} \]

\[ 2.5 \]

**Feature:** \( \phi_{wti} \): Word type \( w \) has hidden state \( t \) at occurrence \( i \)

**Constraint:**
\[
\min_{c_{wt}} \sum_{wt} c_{wt} \quad s.t. \quad E_q(z)[\phi_{wti}] \leq c_{wt}
\]
PoS Induction Results
[Graça et al. 09]

Distribution of word ambiguity

updated slides: http://sideinfo.wikii.com
PoS Induction Results
[Graça et al. 09]

Distribution of word ambiguity

Average ambiguity difference

updated slides: http://sideinfo.wikki.com
PoS Induction Evaluation

[Graça et al. 09]
PoS Induction
Evaluation
[Graça et al. 09]

Mapping from state to pos
PoS Induction Evaluation
[Graça et al. 09]

Mapping from state to pos

Evaluate accuracy

6.5 % Average Improvement

updated slides: http://sideinfo.wikkii.com
PoS Induction Results

[Graça et al. 09]
PoS Induction Results

[Graça et al. 09]

EM - Training

PREP hire (3.4)

TO merge (2.8)

V run (5.8)

N china (7.6)

DET u.s. (7.9)

ADJ

Edges: between word tags at decode time

Very high tag ambiguity

updated slides: http://sideinfo.wikkii.com
PoS Induction Results
[Graça et al. 09]

EM - Training

PREP → hire (3.4)
TO → merge (2.8)
V → run (5.8)
N → china (7.6)
DET → u.s. (7.9)

PR - Training

PREP → hire (1.0)
TO → merge (1.1)
V → run (2.5)
N → china (1.1)
DET → u.s. (1.9)

updated slides: http://sideinfo.wikki.com
Dependency Parsing

DMV Model

[Klein and Manning 04]

Regularization

N creates V sparse ADJ grammars

\[ p_{\theta}(x, y) = \]
Dependency Parsing
DMV Model

[Klein and Manning 04]

\[ p_\theta(x, y) = \theta_{\text{root}}(V) \]

Regularization

\[ \rightarrow \]

creates

sparse

grammars

\[ N \mid V, \text{right}, \text{false} \]

\[ N \mid V, \text{left}, \text{false} \]

\[ N \mid V, \text{right} \]

\[ \theta_{\text{stop}}(\text{nostop} | V, \text{right}, \text{false}) \]

\[ \theta_{\text{stop}}(\text{stop} | V, \text{right}, \text{true}) \]

\[ \theta_{\text{child}}(N \mid V, \text{right}) \]
Dependency Parsing
DMV Model
[Klein and Manning 04]

\[ p_{\theta}(x, y) = \theta_{\text{root}}(V) \]
\[ \cdot \theta_{\text{stop}}(nostop|V, right, false) \]

updated slides: http://sideinfo.wikki.com
Dependency Parsing
DMV Model
[Klein and Manning 04]

\[ p_{\theta}(x, y) = \theta_{\text{root}}(V) \]
\[ \cdot \theta_{\text{stop}}(\text{nostop}|V, \text{right}, \text{false}) \cdot \theta_{\text{child}}(N|V, \text{right}) \]
\[ \cdot \theta_{\text{stop}}(\text{stop}|V, \text{right}, \text{true}) \]
Dependency Parsing
DMV Model
[Klein and Manning 04]

\[
p_\theta(x, y) = \theta_{\text{root}}(V) \\
\quad \cdot \theta_{\text{stop}}(\text{nostop}|V,\text{right},\text{false}) \cdot \theta_{\text{child}}(N|V,\text{right}) \\
\quad \cdot \theta_{\text{stop}}(\text{stop}|V,\text{right},\text{true}) \cdot \theta_{\text{stop}}(\text{nostop}|V,\text{left},\text{false}) \\
\]
Dependency Parsing
DMV Model
[Klein and Manning 04]

\[ p_\theta(x, y) = \theta_{\text{root}}(V) \]
\[ \cdot \theta_{\text{stop}}(\text{nostop}|V, \text{right}, \text{false}) \cdot \theta_{\text{child}}(N|V, \text{right}) \]
\[ \cdot \theta_{\text{stop}}(\text{stop}|V, \text{right}, \text{true}) \cdot \theta_{\text{stop}}(\text{nostop}|V, \text{left}, \text{false}) \cdot \theta_{\text{child}}(N|V, \text{left}) \]
Dependency Parsing
DMV Model
[Klein and Manning 04]

\[ p_\theta(x, y) = \theta_{\text{root}}(V) \]
\[ \cdot \theta_{\text{stop}}(\text{nostop}|V, \text{right}, \text{false}) \cdot \theta_{\text{child}}(N|V, \text{right}) \]
\[ \cdot \theta_{\text{stop}}(\text{stop}|V, \text{right}, \text{true}) \cdot \theta_{\text{stop}}(\text{nostop}|V, \text{left}, \text{false}) \cdot \theta_{\text{child}}(N|V, \text{left}) \]

updated slides: http://sideinfo.wikki.com
Dependency Parsing

- **Minimize number of child/parent relations** [Gillenwater et al. 11]
- **Transfer edges between languages** [Ganchev et al. 09]
- **Use linguistic rules** [Druck et al. 09] [Naseem et al. 10]
Dependency Parsing
Minimize child/parent relations

[Gillenwater et al. 11]
Dependency Parsing
Minimize child/parent relations
[Gillenwater et al. 11]

• ML learns very ambiguous grammars
Dependency Parsing
Minimize child/parent relations
[Gillenwater et al. 11]

• ML learns very ambiguous grammars
  • all productions have some probability
Dependency Parsing
Minimize child/parent relations
[Gillenwater et al. 11]

• ML learns very ambiguous grammars
  • all productions have some probability
  • constrain the number of possible productions
Dependency Parsing
L1LMax over parent/child relations

[ Gillenwater et al. 11 ]

Updated slides: http://sideinfo.wikkii.com
Dependency Parsing
L1LMax over parent/child relations

[Gillenwater et al. 11]
Dependency Parsing
L1LMax over parent/child relations
[Gillenwater et al. 11]

updated slides: http://sideinfo.wikkii.com
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

El sector avícola tiene características muy específicas.

- **Induce Grammar for Spanish -- No resources**
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

El sector avícola tiene características muy específicas.

The poultry sector has very specific characteristics.

- Induce Grammar for Spanish -- No resources
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

El sector avícola tiene características muy específicas.

The poultry sector has very specific characteristics.

- Induce Grammar for Spanish -- No resources
- Have grammar for English
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

El sector avícola tiene características muy específicas.

The poultry sector has very specific characteristics.

- **Induce Grammar for Spanish -- No resources**
- **Have grammar for English**
- **Have parallel text**
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

El sector avícola tiene características muy específicas.

The poultry sector has very specific characteristics.

updated slides: http://sideinfo.wikkii.com
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

El sector avícola tiene características muy específicas.

The poultry sector has very specific characteristics.

• Transfer edges between languages
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

El sector avícola tiene características muy específicas.

The poultry sector has very specific characteristics.

- Transfer edges between languages
- Pick an edge in English

updated slides: http://sideinfo.wikii.com
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

El sector avícola tiene características muy específicas.

The poultry sector has very specific characteristics.

- **Transfer edges between languages**
- **Pick an edge in English**
- **See if child aligns**
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

- **Transfer edges between languages**
  - **Pick an edge in English**
  - **See if child aligns**
  - **See if parent aligns**
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

- Transfer edges between languages
  - Pick an edge in English
  - See if child aligns
  - See if parent aligns
  - Transfer edge

El sector avícola tiene características muy específicas .
The poultry sector has very specific characteristics .
Dependency Parsing

Transfer edges

[Ganchev et al. 09]

• Transfer edges between languages
  • Pick an edge in English
  • See if child aligns
  • See if parent aligns
  • Transfer edge

\[ C_x : \text{Set of transferred edges} \]

updated slides: http://sideinfo.wikki.com
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

El sector avícola tiene características muy específicas.
The poultry sector has very specific characteristics.

- Not all edges are transferred
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

El sector avícola tiene características muy específicas.
The poultry sector has very specific characteristics.

- Not all edges are transferred
- Not all transferred edges are correct
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

• Not all edges are transferred
• Not all transferred edges are correct
• Robust Transfer:

updated slides: http://sideinfo.wikkii.com
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

- Not all edges are transferred
- Not all transferred edges are correct
- Robust Transfer:
  - n% of the transferred edges should be present in the parse

El sector avícola tiene características muy específicas.
The poultry sector has very specific characteristics.
Dependency Parsing

Transfer edges

[Ganchev et al. 09]

\[ C_X : \text{Set of transferred edges} \]

El sector avícola tiene características muy específicas.
The poultry sector has very specific characteristics.
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

$C_x$ : Set of transferred edges

Feature: $\phi(x, y) = \#y \in y \& y \in C_x$
Dependency Parsing
Transfer edges

[Ganchev et al. 09]

\[ C_x : \text{Set of transferred edges} \]

**Feature:** \[ \phi(x, y) = \#y \in y \& y \in C_x \]

**Constraint:** \[ E_q[\phi(x, y)] = \frac{1}{|C_x|} \sum_{y \in C_x} q(y|x) > b \]
Dependency Parsing
Transfer edges
[Ganchev et al. 09]

Accuracy

DMV  PR-Transfer

updated slides: http://sideinfo.wikkii.com
Dependency Parsing
Linguistic Rules
[Naseem et al. 10]

What if no parallel data?

updated slides: http://sideinfo.wikkii.com
Dependency Parsing
Linguistic Rules
[Naseem et al. 10]

What if no parallel data?

Instead small number of universal rules:
What if no parallel data?

Instead small number of universal rules:

<table>
<thead>
<tr>
<th>Rule</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root → Auxiliary</td>
<td>Noun → Adjective</td>
</tr>
<tr>
<td>Root → Verb</td>
<td>Noun → Article</td>
</tr>
<tr>
<td>Verb → Noun</td>
<td>Noun → Noun</td>
</tr>
<tr>
<td>Verb → Pronoun</td>
<td>Noun → Numeral</td>
</tr>
<tr>
<td>Verb → Adverb</td>
<td>Preposition → Noun</td>
</tr>
<tr>
<td>Verb → Verb</td>
<td>Adjective → Adverb</td>
</tr>
<tr>
<td>Auxiliary → Verb</td>
<td></td>
</tr>
</tbody>
</table>
Abstract

We present an approach to grammar induction that utilizes syntactic universals to improve dependency parsing across a range of languages. Our method uses a single set of manually specified language-independent rules that identify syntactic dependencies between pairs of syntactic categories that commonly occur across languages. During inference of the probabilistic model, we use posterior expectation constraints to require that a minimum proportion of the dependencies we infer be instances of these rules. We also automatically refine the syntactic categories given in our coarsely tagged input. Across six languages, our approach outperforms state-of-the-art unsupervised methods by a significant margin.

1 Introduction

Despite surface differences, human languages exhibit striking similarities in many fundamental aspects of syntactic structure. These structural correspondences, referred to as syntactic universals, have been extensively studied in linguistics and underlie many approaches in multilingual parsing. In fact, much recent work has demonstrated that learning cross-lingual correspondences from corpus data greatly reduces the ambiguity inherent in syntactic analysis.

Table 5: The manually specified universal dependency rules used in our experiments. These rules specify head-dependent relationships between coarse, unsplit syntactic categories. An explanation of the ruleset is provided in Section 2.

\[ \mathcal{C}_x \] : All edges in grammar

\[ \mathcal{C}_x \] : All edges in grammar
Dependency Parsing
Linguistic Rules
[Naseem et al. 10]

<table>
<thead>
<tr>
<th>Rule</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root $\rightarrow$ Auxiliary</td>
<td>Noun $\rightarrow$ Adjective</td>
</tr>
<tr>
<td>Root $\rightarrow$ Verb</td>
<td>Noun $\rightarrow$ Article</td>
</tr>
<tr>
<td>Verb $\rightarrow$ Noun</td>
<td>Noun $\rightarrow$ Noun</td>
</tr>
<tr>
<td>Verb $\rightarrow$ Pronoun</td>
<td>Noun $\rightarrow$ Numeral</td>
</tr>
<tr>
<td>Verb $\rightarrow$ Adverb</td>
<td>Preposition $\rightarrow$ Noun</td>
</tr>
<tr>
<td>Verb $\rightarrow$ Verb</td>
<td>Adjective $\rightarrow$ Adverb</td>
</tr>
<tr>
<td>Auxiliary $\rightarrow$ Verb</td>
<td></td>
</tr>
</tbody>
</table>
Dependency Parsing
Linguistic Rules

[Naseem et al. 10]

<table>
<thead>
<tr>
<th>Rule</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Root → Auxiliary</td>
<td>Noun → Adjective</td>
</tr>
<tr>
<td>Root → Verb</td>
<td>Noun → Article</td>
</tr>
<tr>
<td>Verb → Noun</td>
<td>Noun → Noun</td>
</tr>
<tr>
<td>Verb → Pronoun</td>
<td>Noun → Numeral</td>
</tr>
<tr>
<td>Verb → Adverb</td>
<td>Preposition → Noun</td>
</tr>
<tr>
<td>Verb → Verb</td>
<td>Adjective → Adverb</td>
</tr>
<tr>
<td>Auxiliary → Verb</td>
<td></td>
</tr>
</tbody>
</table>

\[ C_x : \text{All edges in grammar} \]

**Feature:** \( \phi(x, y) = \#y \in y \land y \in C_x \)

**Constraint:** \( E_q[\phi(x, y)] = \frac{1}{|y|} \sum_{y \in C_x} q(y|x) > b \)

updated slides: http://sideinfo.wikki.com
Dependency Parsing
Linguistic Rules
[Naseem et al. 10]
Dependency Parsing: Applications using Other Models

• **Tree CRF** [Druck et al. 09]
  • Use universal rules

• **MST Parser** [Ganchev et al. 09]
  • Transfer edges
Information Extraction
Other applications

• **Max-Ent** [Mann et al. 07]
  • Constraints on label marginals

• **CRF** [Druck et al. 09]
  • Actively labeled features

• **Alignment CRF** [Bellare et al. 09]
  • Labeled features

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Information Extraction
Other applications

- **Semi-Markov CRF** [Singh et al. 10]
  - Labeled gazetteers
- **HMM** [Druck et al. 10]
  - Constraints derived from labeled data
Other Applications

- **Multi view learning:** [Ganchev et al. 08]
- **Relation extraction:** [Chen et al. 11]
- ......

updated slides: http://sideinfo.wikii.com
Implementation
Available Software
Off-the-Shelf Tools: MALLET
http://mallet.cs.umass.edu
Off-the-Shelf Tools: MALLET
http://mallet.cs.umass.edu

• *off-the-shelf* support for *labeled features*
Off-the-Shelf Tools: MALLE

http://mallet.cs.umass.edu

- off-the-shelf support for labeled features
- models: MaxEnt Classifier, Linear Chain CRF (one and two label constraints)
Off-the-Shelf Tools: MALLET

http://mallet.cs.umass.edu

- **off-the-shelf** support for **labeled features**
- **models**: MaxEnt Classifier, Linear Chain CRF (one and two label constraints)
- **methods**: GE and PR
Off-the-Shelf Tools: MALLET
http://mallet.cs.umass.edu

- off-the-shelf support for labeled features
- models: MaxEnt Classifier, Linear Chain CRF (one and two label constraints)
- methods: GE and PR
- constraints on label distributions for input features
Off-the-Shelf Tools: MALLET
http://mallet.cs.umass.edu

- **off-the-shelf** support for **labeled features**
- **models**: MaxEnt Classifier, Linear Chain CRF (one and two label constraints)
- **methods**: GE and PR
- **constraints** on label distributions for input features
- **GE penalties**: KL divergence, $L^2_2$ (+ soft inequalities)

updated slides: http://sideinfo.wikki.com
Off-the-Shelf Tools: MALLET
http://mallet.cs.umass.edu

• **off-the-shelf** support for **labeled features**

• **models**: *MaxEnt Classifier*, *Linear Chain CRF* (one and two label constraints)

• **methods**: *GE* and *PR*

• **constraints** on label distributions for input features

• **GE penalties**: KL divergence, $L^2_2$ (+ soft inequalities)

• **PR penalties**: $L^2_2$ (+ soft inequalities)
Off-the-Shelf Tools: MALLET
http://mallet.cs.umass.edu

• off-the-shelf support for labeled features
• models: MaxEnt Classifier, Linear Chain CRF (one and two label constraints)
• methods: GE and PR
• constraints on label distributions for input features
• GE penalties: KL divergence, $L_2^2$ (+ soft inequalities)
• PR penalties: $L_2^2$ (+ soft inequalities)
• in development: Tree CRF, $L_1$ and other penalties

updated slides: http://sideinfo.wikki.com
Off-the-Shelf Tools: MALLET
http://mallet.cs.umass.edu

updated slides: http://sideinfo.wikii.com
Off-the-Shelf Tools: MALLET
http://mallet.cs.umass.edu

- **import data** in SVMLight-like or CoNLL03-like formats

<table>
<thead>
<tr>
<th>positive interesting:2 film:1 ...</th>
<th>U.N.       NNP  B-NP  B-ORG</th>
</tr>
</thead>
<tbody>
<tr>
<td>negative tired:1 sequel:1 ...</td>
<td>official     NN   I-NP  O</td>
</tr>
<tr>
<td>positive best:1 recommend:2 ...</td>
<td>heads        VBZ  B-VP  O</td>
</tr>
</tbody>
</table>

updated slides: http://sideinfo.wikkii.com
Off-the-Shelf Tools: MALLET
http://mallet.cs.umass.edu

• **import data** in SVMLight-like or CoNLL03-like formats

| positive interesting:2 film:1 ... | U.N.       NNP    B-NP    B-ORG |
| negative tired:1 sequel:1 ...    | official   NN     I-NP    O     |
| positive best:1 recommend:2 ...  | heads      VBZ    B-VP    O     |

• **import constraints** in a simple text format:

| tired negative:0.8 positive:0.2 | U.N.  B-ORG:0.7,0.9 |
| best positive:0.9 negative:0.1  | B-VP   0:0.95,1    |
Off-the-Shelf Tools: MALLET

http://mallet.cs.umass.edu

- **import data** in SVMLight-like or CoNLL03-like formats

  positive interesting:2 film:1 ...
negative tired:1 sequel:1 ...
positive best:1 recommend:2 ...

  U.N.       NNP  B-NP  B-ORG
official   NN   I-NP  O
heads      VBZ  B-VP  O

- **import constraints** in a simple text format:

  tired negative:0.8 positive:0.2
best positive:0.9 negative:0.1

  U.N. B-ORG:0.7,0.9
  B-VP 0:0.95,1

- easily **specify method options** (i.e. SimpleTagger):

  java cc.mallet.fst.semi_supervised.tui.SimpleTaggerWithConstraints \
  --train true --test lab --penalty 12 --learning ge \
unlabeled.txt test.txt constraints.txt

updated slides: http://sideinfo.wikkii.com
API for New GE Constraints: MALLE
http://mallet.cs.umass.edu
API for New GE Constraints: MALLET

http://mallet.cs.umass.edu

- *Java Interfaces* for implementing *new* GE constraints
API for New GE Constraints: MALLET
http://mallet.cs.umass.edu

- *Java Interfaces* for implementing *new* GE constraints
- covariance computation implemented (MaxEnt, CRF)
API for New GE Constraints: MALLET
http://mallet.cs.umass.edu

• *Java Interfaces* for implementing new GE constraints
• covariance computation implemented (MaxEnt, CRF)
• **primarily need to write code to:**
  • compute constraint features
  • compute penalty and penalty-specific part of the gradient
API for New GE Constraints: MALLET
http://mallet.cs.umass.edu

- Java Interfaces for implementing new GE constraints
- covariance computation implemented (MaxEnt, CRF)
- primarily need to write code to:
  - compute constraint features
  - compute penalty and penalty-specific part of the gradient
- restriction: constraints must factor with model
- restriction: penalty should be differentiable
API for New PR Constraints: MALLET

http://mallet.cs.umass.edu
API for New PR Constraints:

MALLET

http://mallet.cs.umass.edu

- Java Interfaces for implementing new PR constraints
API for New PR Constraints: MALLET

http://mallet.cs.umass.edu

• *Java Interfaces* for implementing **new** PR constraints

• inference algorithms implemented (MaxEnt, CRF)
API for New PR Constraints: MALLET
http://mallet.cs.umass.edu

- *Java Interfaces* for implementing *new* PR constraints
- inference algorithms implemented (MaxEnt, CRF)
- **primarily need to write code to:**
  - *compute constraint features*
  - *compute penalty and penalty-specific part of the gradient for the modified E-step*
API for New PR Constraints: MALLET
http://mallet.cs.umass.edu

• Java Interfaces for implementing new PR constraints
• inference algorithms implemented (MaxEnt, CRF)

• primarily need to write code to:
  • compute constraint features
  • compute penalty and penalty-specific part of the gradient for the modified E-step

• restriction: constraints must factor with model
Off-the-Shelf Tools & API:
PR Toolkit
http://code.google.com/p/pr-toolkit/

updated slides: http://sideinfo.wikii.com
Off-the-Shelf Tools & API: PR Toolkit
http://code.google.com/p/pr-toolkit/

- off-the-shelf support for PR
Off-the-Shelf Tools & API: PR Toolkit
http://code.google.com/p/pr-toolkit/

- off-the-shelf support for **PR**

- **models:**
  - MaxEnt Classifier, HMM, DMV
Off-the-Shelf Tools & API: PR Toolkit

http://code.google.com/p/pr-toolkit/

• off-the-shelf support for PR

• models:
  • MaxEnt Classifier, HMM, DMV

• applications:
  • Word Alignment, Pos Induction, Grammar Induction
Off-the-Shelf Tools & API: PR Toolkit

http://code.google.com/p/pr-toolkit/

- off-the-shelf support for PR

- **models:**
  - MaxEnt Classifier, HMM, DMV

- **applications:**
  - Word Alignment, Pos Induction, Grammar Induction

- **constraints:** posterior sparsity, bijectivity, agreement
Off-the-Shelf Tools & API: PR Toolkit

http://code.google.com/p/pr-toolkit/

- off-the-shelf support for PR
- models:
  - MaxEnt Classifier, HMM, DMV
- applications:
  - Word Alignment, Pos Induction, Grammar Induction
- constraints: posterior sparsity, bijectivity, agreement
- No command line mode
- Smaller support base
Other Software Packages
Other Software Packages

• **Learning Based Java:**
  • [http://cogcomp.cs.illinois.edu/page/software_view/11](http://cogcomp.cs.illinois.edu/page/software_view/11)
  • support for Constrained Conditional Models

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
Other Software Packages

- **Learning Based Java:**
  - [http://cogcomp.cs.illinois.edu/page/software_view/11](http://cogcomp.cs.illinois.edu/page/software_view/11)
  - support for Constrained Conditional Models

- **Factorie:**
  - support for GE and PR in development
Implementing from Scratch
GE Implementation Example: per-corpus constraints, $L^2_2$ penalty
GE Implementation Example: per-corpus constraints, $L_2^2$ penalty

Compute value and gradient for numerical optimizer:
GE Implementation Example: per-corpus constraints, $L^2_2$ penalty

Compute value and gradient for numerical optimizer:

```c
// compute constraint feature expectations
foreach x : E_{\theta}[\phi] += E_{p_\theta(y|x)}[\phi(x, y)]
```
GE Implementation Example: per-corpus constraints, $L^2_2$ penalty

Compute value and gradient for numerical optimizer:

```plaintext
// compute constraint feature expectations
foreach x : Eθ[φ] += Epθ(y|x)[φ(x, y)]

// compute value
value = −∥b − Eθ[φ]∥^2_2 − \frac{1}{2σ^2} ∥θ∥^2_2
```

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
GE Implementation Example:
per-corpus constraints, $L_2^2$ penalty

Compute value and gradient for numerical optimizer:

// compute constraint feature expectations
foreach $x : E_\theta[\phi] += E_{p_\theta(y|x)}[\phi(x,y)]$

// compute value
value = $-\|b - E_\theta[\phi]\|_2^2 - \frac{1}{2\sigma^2} \|\theta\|_2^2$

// compute gradient
GE Implementation Example: per-corpus constraints, $L^2_2$ penalty

Compute value and gradient for numerical optimizer:

```plaintext
// compute constraint feature expectations
foreach x : E_θ[φ] += E_{pθ(y|x)}[φ(x, y)]

// compute value
value = −∥b − E_θ[φ]∥_2^2 − \frac{1}{2\sigma^2} ∥θ∥_2^2

// compute gradient
gradient = −\frac{1}{\sigma^2} θ
```

updated slides: http://sideinfo.wikkii.com
GE Implementation Example:
per-corpus constraints, $L_2^2$ penalty

Compute value and gradient for numerical optimizer:

// compute constraint feature expectations
foreach $x : E_{\theta}[\phi] += E_{p_{\theta}(y|x)}[\phi(x, y)]$

// compute value
value = $-\|b - E_{\theta}[\phi]\|_2^2 - \frac{1}{2\sigma^2} \|\theta\|_2^2$

// compute gradient
gradient = $-\frac{1}{\sigma^2} \theta$

foreach $x : // for each example
GE Implementation Example: per-corpus constraints, $L^2_2$ penalty

Compute value and gradient for numerical optimizer:

```c
// compute constraint feature expectations
foreach x : E_\theta[\phi] += E_{p_\theta(y|x)}[\phi(x, y)]

// compute value
value = -\|b - E_\theta[\phi]\|^2_2 - \frac{1}{2\sigma^2} \|\theta\|^2_2

// compute gradient
gradient = -\frac{1}{\sigma^2} \theta

foreach x : // for each example
  gradient += 2(b - E_\theta[\phi])^T \text{Cov}_{p_\theta(y|x)}(\phi(x, y), f(x, y))
```

updated slides: http://sideinfo.wikki.com
GE Implementation Example: Intuition
GE Implementation Example: Intuition

- gradient for $\mathbf{x}$: $2(\mathbf{b} - \mathbb{E}_\theta[\phi])^T \text{Cov}_{p_{\theta}(y|x)}(\phi(x, y), f(x, y))$
GE Implementation Example: Intuition

• gradient for $\mathbf{x}$: $2(\mathbf{b} - \mathbb{E}_\theta[\phi])^T \text{Cov}_{p_\theta}(\mathbf{y}|\mathbf{x})(\phi(\mathbf{x}, \mathbf{y}), f(\mathbf{x}, \mathbf{y}))$

• example: model expectation $<$ target expectation
GE Implementation Example: Intuition

- gradient for $x$: $2(b - E_\theta[\phi])^T \text{Cov}_{p_\theta(y|x)}(\phi(x, y), f(x, y))$

- **example**: model expectation < target expectation
GE Implementation Example: Intuition

- gradient for $\mathbf{x}$: $2(b - E_{\theta}[\phi])^T \text{Cov}_{\theta}(y|x)(\phi(x, y), f(x, y))$
- example: model expectation $<$ target expectation

which parameters should be increased to increase the model expectation?
GE Implementation Example: Intuition

- gradient for $\mathbf{x}$: $2(b - \mathbb{E}_\theta[\phi])^T \text{Cov}_{p_\theta(y|x)}(\phi(x, y), f(x, y))$

- example: model expectation < target expectation

  which parameters should be increased to increase the model expectation?

  (which parameters have large gradient)

updated slides: http://sideinfo.wikkii.com
GE Implementation Example: Intuition

• gradient for \( \mathbf{x} \):
  \[
  2(\mathbf{b} - \mathbb{E}_\theta[\phi])^T \text{Cov}_{\theta}(\mathbf{y}|\mathbf{x})(\phi(\mathbf{x}, \mathbf{y}), f(\mathbf{x}, \mathbf{y}))
  \]

• example: model expectation < target expectation

  which parameters should be increased to increase the model expectation?

  (which parameters have large gradient)

• parameters for model features with highest positive covariance with constraint features
GE Implementation Example: Intuition

• gradient for $\mathbf{x}$: $2(\mathbf{b} - \mathbb{E}_{\theta}[\phi])^T \text{Cov}_{p_{\theta}(y|x)}(\phi(x, y), f(x, y))$

• example: model expectation < target expectation

  which parameters should be increased to increase the model expectation?

  (which parameters have large gradient)

• parameters for model features with highest positive covariance with constraint features

• magnitude depends on covariance, difference from target
GE Implementation Example: Intuition

- gradient for \( x \): \[ 2(b - E_\theta[\phi])^T \text{Cov}_{p_\theta(y|x)}(\phi(x, y), f(x, y)) \]
- example: model expectation < target expectation
  - which parameters should be increased to increase the model expectation? (which parameters have large gradient)
- parameters for model features with highest positive covariance with constraint features
- magnitude depends on covariance, difference from target

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
GE Implementation Example: Intuition

- gradient for $\mathbf{x}$: $2(\mathbf{b} - \mathbb{E}_\theta[\phi])^\top \text{Cov}_{p_{\theta}(\mathbf{y}|\mathbf{x})}(\phi(\mathbf{x}, \mathbf{y}), f(\mathbf{x}, \mathbf{y}))$

- **example**: model expectation $<$ target expectation

  which parameters should be increased to increase the model expectation?

  (which parameters have large gradient)

- parameters for *model features* with highest positive *covariance* with *constraint features*

- magnitude depends on *covariance*, *difference from target*

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
GE Implementation Example: Computing Gradient

- **trick:** never need to compute / store a matrix

\[ 2(b - E_\theta[\phi])^T \left( E_\theta[\phi f^T] - E_\theta[\phi] E_\theta[f^T] \right) \]
GE Implementation Example: Computing Gradient

- **trick:** never need to compute / store a matrix

\[
2(b - E_\theta[\phi])^T \left( E_\theta[\phi f^T] - E_\theta[\phi] E_\theta[f^T] \right)
\]

\[
= E_\theta \left[ 2(b - E_\theta[\phi])^T \phi f^T \right] - E_\theta \left[ 2(b - E_\theta[\phi])^T \phi \right] E_\theta[f^T]
\]
GE Implementation Example: Computing Gradient

- **trick:** never need to compute / store a matrix

\[
2(b - \mathbb{E}_\theta[\phi])^T(E_\theta[\phi]f^T - \mathbb{E}_\theta[\phi]E_\theta[f^T])
\]

\[
= \mathbb{E}_\theta[2(b - \mathbb{E}_\theta[\phi])^T \phi f^T] - \mathbb{E}_\theta[2(b - \mathbb{E}_\theta[\phi])^T \phi E_\theta[f^T]]
\]
GE Implementation Example: Computing Gradient

- **trick:** never need to compute / store a matrix

\[
2(b - E_\theta[\phi])^T\left(E_\theta[\phi f^T] - E_\theta[\phi] E_\theta[f^T]\right)
\]

\[
= E_\theta[2(b - E_\theta[\phi])^T \phi f^T] - E_\theta[2(b - E_\theta[\phi])^T \phi] E_\theta[f^T]
\]
GE Implementation Example: Computing Gradient

- **trick**: compute \(\text{Cov}\) with composite constraint feature

- \(\phi_c(x, y) = 2(b - E_\theta[\phi])^T\phi(x, y)\)

- **result**: compute/store vectors of size \(\text{dim}(f)\) (never a matrix)
GE Implementation Example: Computing Gradient

- **trick:** compute $\text{Cov}$ with composite constraint feature

  $$\phi_c(x, y) = 2(b - E_\theta[\phi])^T \phi(x, y)$$

- **result:** compute/store vectors of size $\dim(f)$ (never a matrix)

- **trick:** if inference can be cast as hypergraph problem, or if the graphical model is a tree

  can use efficient semiring algorithms to compute $\text{Cov}$
  [Li & Eisner 09] [Pauls et al. 09]

- **result (w. both):** same time complexity as standard inference

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)
PR Implementation Example:
Word Alignment, Bijectivity

updated slides: http://sideinfo.wikii.com
PR Implementation Example: Word Alignment, Bijectivity

• **Learning**: EM, PR
  
  • `void eStep(counts, lattices);`
  • `void mStep(counts);`
  • `lattice constraint.project(lattice);`
PR Implementation Example: Word Alignment, Bijectivity

- **Learning**: EM, PR
  - `void eStep(counts, lattices);`
  - `void mStep(counts);`
  - `lattice constraint.project(lattice);`

- **Model**: HMM
  - `lattice computePosteriors(lattice);`
  - `void addCount(lattice, counts);`
  - `void updateParameters(counts);`
PR Implementation Example: Word Alignment, Bijectivity

- **Learning**: EM, PR
  - void eStep(counts, lattices);
  - void mStep(counts);
  - lattice constraint.project(lattice);

- **Model**: HMM
  - lattice computePosteriors(lattice);
  - void addCount(lattice, counts);
  - void updateParameters(counts);

- **Constraints**: Bijectivity
  - lattice project(lattice);

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
PR Implementation Example: EM

class EM {
    model;

    void em(n) {
        lattices = model.getLattices();
        counts = model.counts();
        for (i=0; i < n; i++) {
            eStep(counts, lattices);
            mStep(counts);
        }
    }

    void eStep(counts, lattices) {
        counts.clear();
        for (l : lattices) {
            l = model.computePosterior(l);
            model.addCount(l, counts);
        }
    }

    void mStep(counts) {
        model.updateParameters(counts);
    }

    ...
}

updated slides: http://sideinfo.wikii.com
class PR {

    model;
    constraint;

    void em(n){
        lattices= model.getLattices();
        counts = model.counts();
        for(i=0; i< n; i++) {
            eStep(counts, lattices);
            mStep(counts);
        }
    }

    void eStep(counts, lattices) {
        counts.clear();
        for(l : lattices){
            l = model.computePosterior(l);
            l = constraint.project(l);
            model.addCount(l,counts);
        }
    }

    void mStep(counts) {
        model.updateParameters(counts);
    }
}

updated slides: http://sideinfo.wikii.com
class HMM {
    obsProb, transProbs, initProbs;

    lattice computerPosteriors(lattice) {
        "Run forward backward"
    }

    void addCount(lattice, counts) {
        "Add posteriors to count table"
    }

    void updateParams(counts) {
        "Normalize counts"
        "Copy counts to params table"
    }

    void getCounts() {
        "return copy of params structures"
    }

    void getLattices() {
        "return structure of all lattices in the corpus"
    }

    .......
}

updated slides: http://sideinfo.wikii.com
PR Implementation Example: Word Alignment, Bijection

• **constraint features:** $\phi(x, y)$

• # target words that align with each source word

• **constraint:** $Q = \{ q : E_q[\phi(x, y)] \leq 1 \}$
PR Implementation Example: Word Alignment, Bijectivity

- **constraint features**: $\phi(x, y)$
- # target words that align with each source word
- **constraint**: $Q = \{ q : E_q[\phi(x, y)] \leq 1 \}$
- **project method**:
PR Implementation Example: Word Alignment, Bijectivity

- **constraint features**: $\phi(x, y)$
- # target words that align with each source word
- **constraint**: $Q = \{ q : E_q[\phi(x, y)] \leq 1 \}$

**project method**:

- **primal** (hard): $D_{KL}(Q|p_\theta) = \arg \min_{q \in Q} D_{KL}(q|p_\theta)$
**PR Implementation Example:** Word Alignment, Bijectivity

- **constraint features:** $\phi(x, y)$
- # target words that align with each source word
- **constraint:** $\mathcal{Q} = \{ q : \mathbf{E}_q[\phi(x, y)] \leq 1 \}$
- **project method:**
  - **primal** (hard): $\mathcal{D}_{KL}(\mathcal{Q}|p_{\theta}) = \arg\min_{q \in \mathcal{Q}} \mathcal{D}_{KL}(q|p_{\theta})$
  - **dual** (easy): $\arg\max_{\lambda \geq 0} -b \cdot \lambda - \log Z(\lambda)$

$$Z(\lambda) = \sum_y p_{\theta}(y|x) \exp(-\lambda \cdot \phi(x, y))$$

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
PR Implementation Example: Bijective Constraints

class BijectiveConstraints {
    model;

    lattice project(lattice){
        obj = BijectiveObj(model, lattice);
        Optimizer.optimize(obj);
        return lattice;
    }
}

class BijectiveObj {
    model, lambda, lattice, b;

    void setParameters(newLambda) {
        lambda = newLambda;
        updateModel();
    }

    void updateModel(){
        lattice = lattice*exp(-lambda);
        lattice = model.computePosteriors(lattice);
    }

    double getObj() {
        obj = -dot(lambda, b);
        obj -= lattice.likelihood;
    }

    double[] getGrad(newLambda) {
        grad = ex(lattice.posteriors) - b;
        return grad;
    }
}

updated slides: http://sideinfo.wikki.com
Which framework should I use?
Which framework should I use?

Open research question...
Which framework should I use?

Open research question...

Really, which framework should I use?
Which framework should I use?

Open research question...

Really, which framework should I use?

Each framework is well-suited to particular applications.
Consider **CODL** When...
Consider **CODL** When...

max inference is *easy*, but computing expectations is *hard*
Consider **CODL** When...

max inference is *easy*, but computing expectations is *hard*

examples:
Consider **CODL** When...

max inference is *easy*, but computing expectations is *hard*

examples:

- **non-projective dependency parsing:**
  - **max**: maximum spanning tree, $O(n^2)$
  - **expectations**: matrix-tree theorem, $O(n^3)$
Consider **CODL** When...

max inference is *easy*, but computing expectations is *hard*

examples:

- **non-projective dependency parsing:**
  - **max:** maximum spanning tree, $O(n^2)$
  - **expectations:** matrix-tree theorem, $O(n^3)$

- tasks where output variables have large cardinality:
  - *storing* expectations may be infeasible
Consider **PR** When...
Consider PR When...

already using EM (modification to use PR is small)
Consider PR When...

already using EM (modification to use PR is small)

• example: directed, generative model; corpus constraints
Consider PR When...

already using EM (modification to use PR is small)

• example: directed, generative model; corpus constraints

• compared to CODL:
  • developing a penalty/inference method may be difficult
Consider PR When...

already using EM (modification to use PR is small)

• example: directed, generative model; corpus constraints

• compared to CODL:
  • developing a penalty/inference method may be difficult

• compared to GE:
  • need to develop gradient-based methods
  • non-parametric model: unclear how to apply GE

updated slides: http://sideinfo.wikii.com
Consider **GE** When...
Consider **GE** When...

already using direct gradient and can compute $\text{Cov}$ efficiently
Consider GE When...

already using direct gradient and can compute $\text{Cov}$ efficiently

• example: linear chain CRF; labeled feature constraints

updated slides: http://sideinfo.wikii.com
Consider **GE** When...

already using direct gradient and can compute $\textbf{Cov}$ efficiently

- example: linear chain CRF; labeled feature constraints

- compared to **CODL**: developing a penalty/inference method may be difficult

updated slides: [http://sideinfo.wikki.com](http://sideinfo.wikki.com)
Consider **GE** When...

already using direct gradient and can compute $\text{Cov}$ efficiently

• example: linear chain CRF; labeled feature constraints

• compared to **CODL**: developing a penalty/inference method may be difficult

• compared to **PR**: in experiments, GE often converges more quickly

updated slides: [http://sideinfo.wikii.com](http://sideinfo.wikii.com)
Consider **GE** When...

- already using direct gradient and can compute **Cov** efficiently
  - example: linear chain CRF; labeled feature constraints
  - compared to **CODL**: developing a penalty/inference method may be difficult
  - compared to **PR**: in experiments, GE often converges more quickly

updated slides: [http://sideinfo.wikkii.com](http://sideinfo.wikkii.com)

![Graph](image_url)
Thanks!

• Learn more at:

http://sideinfo.wikkii.com