Discriminative Modeling Topics

April 2, 2013
Menu du Jour

• MaxEnt phrase reordering
  (Xiong et al., 2006)

• Discriminative lexicon models
  (Mauser et al., 2009)

• Translation as CRFs
  (Blunsom et al., 2008)
SCFGs: A problem

A → bruja | witch  0.2
A → verde | green  0.2
A → A A | 1 2  0.6
A → A A | 2 I  0.2
SCFGs: A problem

A → bruja | witch 0.2
A → verde | green 0.2
A → A A | 1 2 0.6
A → A A | 2 1 0.2

0.6
SCFGs: A problem

A → bruja | witch 0.2
A → verde | green 0.2
A → A A | 1 2 0.6
A → A A | 2 1 0.2

0.6x0.2
SCFGs: A problem

A → bruja | witch 0.2
A → verde | green 0.2
A → A A | 1 2 0.6
A → A A | 2 1 0.2

0.6 x 0.2 x 0.2 = 0.024
SCFGs: A problem

\[
\begin{align*}
A & \rightarrow \text{bruja} \mid \text{witch} \quad 0.2 \\
A & \rightarrow \text{verde} \mid \text{green} \quad 0.2 \\
A & \rightarrow AA \mid 1 2 \quad 0.6 \\
A & \rightarrow AA \mid 2 1 \quad 0.2
\end{align*}
\]

\[
0.6 \times 0.2 \times 0.2 = 0.024 \quad \text{witch} \quad \text{green}
\]
SCFGs: A problem

A → bruja | witch 0.2
A → verde | green 0.2
A → A A | 1 2 0.6
A → A A | 2 1 0.2

\[0.2 \times 0.2 \times 0.2 = 0.008\]
SCFGs: A problem

A → bruja | witch 0.2
A → verde | green 0.2
A → A A | 1 2 0.6
A → A A | 2 1 0.2

0.2×0.2×0.2 = 0.008

Context-free rules apply independent of context.
Some solutions

- More sophisticated grammars
  \[ A \rightarrow A A \mid 2 \]
Some solutions

- More sophisticated grammars

\[
A \rightarrow AA | 2 \ 1
\]

\[
NP \rightarrow NN \ JJ | 2 \ 1
\]
Some solutions

• More sophisticated grammars

\[ A \rightarrow AA | 2 \]

\[ NP \rightarrow NN JJ | 2 \]

What are the problems with this?
Some solutions

- More sophisticated grammars
  \[ A \rightarrow AA | 2 \]
  \[ NP \rightarrow NN JJ | 2 \]

- Discriminative “parsing”
  \[ p(A = \{\leftarrow, \rightarrow\} | f, A_{\text{left}}, A_{\text{right}}) \]

bruja verde
Some solutions

• More sophisticated grammars

\[ A \rightarrow AA | 2 \]
\[ NP \rightarrow NN JJ | 2 \]

• Discriminative “parsing”

\[ p(A = \{\leftarrow, \rightarrow\} | f, A_{\text{left}}, A_{\text{right}}) \]

Source
Left child
Right child

bruja
derde
Key Insight

- PCFGs are generative models of text (parallel text)
- In translation, the text is given: use discriminative models
- Xiong et al. propose a very simple approach:
  - Standard translation model (phrase based)
  - Standard (uniform) segmentation model
  - Standard n-gram language model
- *Innovation*: every time you form a constituent, predict whether it should be *monotone* or *inverted*
Again, we reduce a major component of translation to **binary classification**.
MaxEnt Model

\[
p(A = \leftarrow | \text{bruja verde}, (0, 1), (1, 2))
\]

This is a lot of conditioning

\[
p(A = \leftarrow, \rightarrow | f, A_{\text{left}}, A_{\text{right}}) = \frac{1}{Z} \exp \mathbf{w}^\top \phi(f, A_{\text{left}}, A_{\text{right}})
\]

Again, we reduce a major component of translation to \textbf{binary classification}. 
Training the Model

• What do we need to train the model?
Training the Model

• What do we need to train the model?

• How do we extract training examples from the training data?
Extract examples from word aligned training data.
Extract examples from word aligned training data.
Extract examples from word aligned training data.
Extract examples from word aligned training data.
Extract examples from word aligned training data.
Extract examples from word aligned training data.
Extract examples from word aligned training data.
## Results

<table>
<thead>
<tr>
<th>Condition</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>monotonic</td>
<td>20.1</td>
</tr>
<tr>
<td>no-model</td>
<td>19.6</td>
</tr>
<tr>
<td>size of constituent</td>
<td>20.9</td>
</tr>
<tr>
<td>MaxEnt</td>
<td>22.2</td>
</tr>
</tbody>
</table>
General Insights

• Decoders make local translation and reordering decisions
  • Standard approach: relative frequency
  • Alternative: using “local classifiers”
  • Challenge: extract (noisy) training instances from the training data
  • Benefits: no decoding required for training these local classifiers

• The source is given: use it!
Questions?
Modeling the Lexicon

bestehen → consist 0.5
       → insist 0.4
       → pass 0.1
Modeling the Lexicon

Es wird morgen eine Prüfung geben

Ob ich bestehen werde?

There’s a test tomorrow.

Will I pass?

bestehen

consist 0.5

insist 0.4

pass 0.1
Modeling the Lexicon

Es wird morgen eine Prüfung geben
Ob ich bestehen werde?

There’s a test tomorrow.
Will I pass?

Whether I shall consist?
Modeling the Lexicon

\[ p_{\text{new model}}(\text{pass} \mid \text{bestehen}, C = \text{Prüfung}) > p(w \mid \text{bestehen}) \]

Es wird morgen eine Prüfung geben

Ob ich bestehen werde?

Goal:

There's a test tomorrow.

Will I pass?

Whether I shall consist?
The DWL

- “Discriminative Word Lexicon”

\[ p(\mathcal{E} \mid \mathcal{F}) \]

Set of English words \quad Set of Foreign words
The DWL

- “Discriminative Word Lexicon”

\[
p(\mathcal{E} \mid \mathcal{F}) = \prod_{e \in \mathcal{E}} p(\text{contains } e \mid \mathcal{F}) \times \prod_{e \in \mathcal{E}^C} (1 - p(\text{contains } e \mid \mathcal{F}))
\]

Model inclusion as conditionally independent binary decisions
Binary Classifiers

• Downside
  • Independence assumptions are harsh

• Upside
  • Training for every word in the vocabulary can be carried out in parallel
Training the Model

• What do we need to train the model?

• How do we extract training examples from the training data?
\[ \Sigma = \{ \text{the, and, of, cat, \ldots, pass, test, \ldots resulting, xylophone} \} \]
\[ \Sigma = \{ \text{the, and, of, cat, \ldots, pass, test, \ldots resulting, xylophone} \} \]

**Sentence pair:**

you will pass the test \hspace{1cm} du wirst die Prüfung bestehen
\[ \Sigma = \{ \text{the, and, of, cat, \ldots, pass, test, \ldots resulting, xylophone} \} \]

Sentence pair:

*you will pass the test*  
*du wirst die Pruefung bestehen*

<table>
<thead>
<tr>
<th>Classifier</th>
<th>(y)</th>
<th>Feature Vector ((x))</th>
</tr>
</thead>
<tbody>
<tr>
<td>pass?</td>
<td>+</td>
<td>du=1 wirst=1 Pruefung=1 besten=1</td>
</tr>
<tr>
<td>will?</td>
<td>+</td>
<td>du=1 wirst=1 Pruefung=1 besten=1</td>
</tr>
<tr>
<td>insist?</td>
<td>-</td>
<td>du=1 wirst=1 Pruefung=1 besten=1</td>
</tr>
<tr>
<td>insist?</td>
<td>-</td>
<td>du=1 wirst=1 Pruefung=1 besten=1</td>
</tr>
<tr>
<td>cat?</td>
<td>-</td>
<td>du=1 wirst=1 Pruefung=1 besten=1</td>
</tr>
<tr>
<td>xylophone?</td>
<td>-</td>
<td>du=1 wirst=1 Pruefung=1 besten=1</td>
</tr>
</tbody>
</table>
\[ \Sigma = \{ \text{the, and, of, cat, \ldots, pass, test, \ldots resulting, xylophone} \} \]

**Sentence pair:**

\textit{you will pass the test} \hspace{1cm} \textit{du wirst die Pruefung bestehen}

<table>
<thead>
<tr>
<th>Classifier</th>
<th>( y )</th>
<th>Feature Vector ((x))</th>
</tr>
</thead>
<tbody>
<tr>
<td>pass?</td>
<td>+</td>
<td>\textmd{(du=1\ wirst=1\ Pruefung=1\ besten=1)}</td>
</tr>
<tr>
<td>will?</td>
<td>+</td>
<td>\textmd{(du=1\ wirst=1\ Pruefung=1\ besten=1)}</td>
</tr>
<tr>
<td>insist?</td>
<td>-</td>
<td>\textmd{(du=1\ wirst=1\ Pruefung=1\ besten=1)}</td>
</tr>
<tr>
<td>insist?</td>
<td>-</td>
<td>\textmd{(du=1\ wirst=1\ Pruefung=1\ besten=1)}</td>
</tr>
<tr>
<td>cat?</td>
<td>-</td>
<td>\textmd{(du=1\ wirst=1\ Pruefung=1\ besten=1)}</td>
</tr>
<tr>
<td>xylophone?</td>
<td>-</td>
<td>\textmd{(du=1\ wirst=1\ Pruefung=1\ besten=1)}</td>
</tr>
</tbody>
</table>

\(O(N \times V)\) training instances
Rescoring with the DWL

- The DWL assigns probabilities to **sets** of words
  - Once a word is used once, subsequent uses are “free”
  - This makes dynamic programming difficult
- A simple strategy: reranking
  - Get $k$-best lists from baseline decoder, compute DWL score on each entry
  - Train a second model (using PRO, MERT, etc.) as if the $k$-best lists were the decoder
  - Search errors are very possible!
### Arabic-English

<table>
<thead>
<tr>
<th>Condition</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>42.0</td>
</tr>
<tr>
<td>+DWL</td>
<td>42.4</td>
</tr>
</tbody>
</table>

### Chinese-English

<table>
<thead>
<tr>
<th>Condition</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>25.3</td>
</tr>
<tr>
<td>+DWL</td>
<td>26.2</td>
</tr>
<tr>
<td>Source</td>
<td>目前，事故抢险组正在紧急恢复通风系统。</td>
</tr>
<tr>
<td>----------</td>
<td>-------------------------------------</td>
</tr>
<tr>
<td>Baseline</td>
<td>at present, the accident and rescue teams are currently emergency recovery ventilation systems.</td>
</tr>
<tr>
<td>DWL</td>
<td>at present, the emergency rescue teams are currently restoring the ventilation system.</td>
</tr>
<tr>
<td>Reference</td>
<td>right now, the accident emergency rescue team is making emergency repair on the ventilation system.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DWL</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>emergency</td>
<td>0.894</td>
</tr>
<tr>
<td>currently</td>
<td>0.330</td>
</tr>
<tr>
<td>current</td>
<td>0.175</td>
</tr>
<tr>
<td>emergencies</td>
<td>0.133</td>
</tr>
<tr>
<td>present</td>
<td>0.133</td>
</tr>
<tr>
<td>accident</td>
<td>0.119</td>
</tr>
<tr>
<td>recovery</td>
<td>0.053</td>
</tr>
<tr>
<td>group</td>
<td>0.046</td>
</tr>
<tr>
<td>dealing</td>
<td>0.042</td>
</tr>
<tr>
<td>ventilation</td>
<td>0.034</td>
</tr>
</tbody>
</table>
Possible Extensions

- Condition on more context than the sentence (e.g., document)
- Model units larger than words (e.g., phrases)
- Model only words/phrases that have ambiguous translations
  - Measure of ambiguity: entropy
Questions?
Translation as CRFs
Translation as CRFs

No linear chains
CRFs on Synchronous Trees

- Model: discriminate “good” parses under an SCFG from bad ones
- Challenges:
  - What is a good parse?
    - One that produces a good translation
  - What is a good translation?
    - The one we see in the training data
  - What about the zillions of ways to derive a sentence pair?
    - Let’s marginalize them
  - What about non-literal translations?
    - Regularize so we don’t memorize “bad” stuff
Which Derivation to Optimize?

the red hat
|
le chapeau
Which Derivation to Optimize?

the red hat
  /
le chapeau

the red hat
  /
le chapeau
Which Derivation to Optimize?
Which Derivation to Optimize?

```
| the red hat | the red hat |
| le chapeau  | le chapeau  |
```

```
| the red hat | the red hat |
| le chapeau  | le chapeau  |
```
Parametric Form

Conditional probability of a derivation

\[ p_{\Lambda}(d, e|f) = \frac{\exp \sum_k \lambda_k H_k(d, e, f)}{Z_{\Lambda}(f)}. \]

Conditional probability of a translation
Features

The features must decompose with the rules:

\[ H_k(d, e, f) = \sum_{r \in d} h_k(f, r, q(r, d)) \]

- Any part of the source may be used
  - Source syntax
  - Morphology
  - Lexical context
  - POS information
Training

\[ \mathcal{L} = \sum_{(e,f) \in \mathcal{D}} \sum_d \log p_\Lambda(e, d \mid f) + \sum_m \frac{\lambda_m^2}{2\sigma^2} \]
Training

\[ \mathcal{L} = \sum_{(e,f) \in D} \sum_d \log p_{\Lambda}(e, d \mid f) + \sum_m \frac{\lambda_m^2}{2\sigma^2} \]

Differentiable:

\[ \frac{\partial \mathcal{L}}{\partial w_i} = \sum_{(e,f) \in D} \mathbb{E}_{p_{\Lambda}(e,d|f)} h_i(e,d,f) - \mathbb{E}_{p_{\Lambda}(d|e,f)} h_i(e,d,f) - \frac{\lambda_i}{\sigma^2} \]
Inference

• How do we compute the following feature expectations?

\[
\frac{\partial L}{\partial w_i} = \sum_{(e,f) \in D} \mathbb{E}_{p_X(e,d|f)} h_i(e,d,f) - \mathbb{E}_{p_X(d|e,f)} h_i(e,d,f) - \frac{\lambda_i}{\sigma^2}
\]
## Effect of Regularization

<table>
<thead>
<tr>
<th>Grammar Rules</th>
<th>ML ($\sigma^2 = \infty$)</th>
<th>MAP ($\sigma^2 = 1$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\langle X \rangle \rightarrow \langle carte, map \rangle$</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>$\langle X \rangle \rightarrow \langle carte, notice \rangle$</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>$\langle X \rangle \rightarrow \langle sur, on \rangle$</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\langle X \rangle \rightarrow \langle la, the \rangle$</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\langle X \rangle \rightarrow \langle table, table \rangle$</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>$\langle X \rangle \rightarrow \langle table, chart \rangle$</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>$\langle X \rangle \rightarrow \langle carte sur, notice on \rangle$</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>$\langle X \rangle \rightarrow \langle carte sur, map on \rangle$</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>$\langle X \rangle \rightarrow \langle sur la, on the \rangle$</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>$\langle X \rangle \rightarrow \langle la table, the table \rangle$</td>
<td>0.0</td>
<td>0.5</td>
</tr>
<tr>
<td>$\langle X \rangle \rightarrow \langle la table, the chart \rangle$</td>
<td>1.0</td>
<td>0.5</td>
</tr>
</tbody>
</table>

**Training data:**
- carte sur la table $\leftrightarrow$ map on the table
- carte sur la table $\leftrightarrow$ notice on the chart
<table>
<thead>
<tr>
<th>Condition</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hiero -LM</td>
<td>28.1</td>
</tr>
<tr>
<td>Hiero +LM</td>
<td>32.0</td>
</tr>
<tr>
<td>CRF - max deriv</td>
<td>25.8</td>
</tr>
<tr>
<td>CRF - max trans</td>
<td>27.7</td>
</tr>
</tbody>
</table>
Questions?