Outline

- 3:20pm Parametric models: posterior regularisation. Desai
- 3:35pm Training models with rich features spaces. Vlad
- 3:50pm Decoding with complex grammars. Adam
- 4:20pm Closing remarks. Phil
- 4:25pm Finish.
Phrase Clustering with Posterior Regularization

CLSP Summer Workshop 2010
SMT Team
Desai Chen
joint work with Trevor Cohn
Outline

- clustering problem
- EM with posterior regularization
- results and future experiments
Phrase clustering

Phrases are defined as contiguous spans aligned with each other

i'll bring you some now.

我这就给您拿一些。

Example from btec
Phrase clustering

Phrases are defined as contiguous spans aligned with each other

"I'll bring you some now."

Example from btec
Phrase clustering

Phrases are defined as contiguous spans aligned with each other

Example from btec
Phrase clustering

Phrases are defined as contiguous spans aligned with each other

i'll bring you some now.

我这就给您拿一些。
Phrase clustering

Contexts are words before or after the phrase:

**target side context**

```
i'll bring you some now.
```

**source side context**

```
我这就给您拿一些。
```
Objective

Put all phrase-context pairs into categories
Objective

Put all phrase-context pairs into categories
Outline

- Where do phrases come from?
- EM with posterior regularization
- Results and future experiment
Expectation-Maximization

- naïve Bayes model for phrase labeling
EM clustering

- naïve Bayes model for phrase labeling
EM clustering

- naïve Bayes model for phrase labeling
EM clustering

- naïve Bayes model for phrase labeling

\[ q(z|p, c) = P_\theta(z|p, c) \]
EM clustering

- naïve Bayes model for phrase labeling

\[ q(z|p,c) = P_\theta(z|p,c) \]

\[ \theta = \text{MLE } q(z|p,c) \]
Problem with EM

- Problem: EM uses as many categories as it wants for each phrase.
- We want to limit the number of categories associated with each phrase.
Sparsity constraints

- Sparsity: Each phrase/context should be labeled with fewer kinds of labels.
Sparsity: Each phrase/context should be labeled with fewer kinds of labels.

Diagram:
- Phrase
- z
- context
- constraint

Constraint:
- Phrases
- Context
- z
Sparsity constraints

Minimize $\sum_{p,z} \max_i P(z|p_i)$
Sparsity constraints

Minimize $\sum_{p,z} \max_i P(z|p_i)$

Phrase: there are

Contexts:
  i understand there are some sightseeing bus tours here , is that right ?

there are only a few seats left in the dress circle .

well , of course there are fine restaurants .

your hotel brochure shows there are some tennis counts at your hotel .
Sparsity constraints

Minimize $\sum_{p,z} \max_i P(z|p_i)$

Phrase: there are

Contexts:
  i understand there are some sightseeing bus tours here, is that right?

there are only a few seats left in the dress circle.

well, of course there are fine restaurants.

your hotel brochure shows there are some tennis counts at your hotel.
Sparsity constraints

Minimize $\sum_{p,z} \max_i P(z|p_i)$

Phrase: there are

Contexts:
i understand _ some
sightseeing

<s> <s> _ only a

of course _ fine
restaurants
brochure shows _
some tennis
Sparsity constraints

Minimize $\sum_{p,z} \max_i P(z | p_i)$

Phrase: there are

Contexts:
i understand _ some sightseeing
<s> <s> _ only a
of course _ fine restaurants
brochure shows _ some tennis
Sparsity constraints

Minimize $\sum_{p,z} \max_i P(z|p_i)$

Phrase: there are

Contexts:
i understand _ some sightseeing
<s> <s> _ only a
of course _ fine restaurants
brochure shows _ some tennis
max P(tag|phrase)
Sparsity constraints

Minimize $\sum_{p,z} \max_i P(z|p_i)$

Phrase: there are

Contexts:
i understand _ some sightseeing
<s> <s> _ only a
of course _ fine restaurants
brochure shows _ some tennis
max $P(\text{tag}|\text{phrase})$
Sparsity constraints

Minimize $\sum_{p,z} \max_i P(z|p_i)$

Phrase: there are

Contexts:
i understand _ some sightseeing

<s> <s> _ only a

of course _ fine restaurants

brochure shows _ some tennis

max $P(\text{tag}|\text{phrase})$
Sparsity constraints

Minimize $\sum_{p,z} \max_i P(z|p_i)$

Phrase: there are

Contexts:
i understand _ some sightseeing
<s> <s> _ only a
of course _ fine restaurants
brochure shows _ some tennis
  max P(tag|phrase)
Sparsity constraints

Minimize $\sum_{p,z} \max_i P(z|p_i)$

Phrase: there are

Contexts:
- i understand _ some sightseeing
- <s> <s> _ only a
- of course _ fine restaurants
- brochure shows _ some tennis
- max P(tag|phrase)

Sum

P(Z=1) | P(Z=2) | P(Z=3) | P(Z=4)

[Color-coded table with varying intensities]
Sparsity constraints

Minimize $\sum_{p,z} \max_i P(z|p_i)$

Phrase: there are

Contexts:
- i understand _ some sightseeing
- <s> <s> _ only a
- of course _ fine restaurants
- brochure shows _ some tennis

max $P(\text{tag}|\text{phrase})$

$\sum \approx 1$
Posterior Regularization

- Follows *Posterior Regularization for Structured Latent Variable Models*, Ganchev et al., 2009
- During E-step, impose constraints on the posterior $q$ to guide the search
Posterior Regularization

- impose constraints on the posterior $q$

$$q(z|p, c) = \arg \min_{q \in Q} KL(q||P_\theta)$$

$\theta = MLE \ q(z|p, c)$
Posterior Regularization

- impose constraints on the posterior $q$

$$q(z|p, c) = \arg\min_{q \in Q} KL(q||P_\theta)$$
Sparsity constraints

Minimize $\sum_{p,z} \max_i P(z|p_i)$

Phrase: like this

Contexts:
i understand _ some sightseeing

Define feature functions:

$\phi_{i,j}(p, z) = \begin{cases} 1 & \text{if } p = i \text{ and } z = j \\ 0 & \text{otherwise} \end{cases}$
Sparsity constraints

Minimize $\sum_{p,z} \max_i P(z|p_i)$

- Soft constraint. Softness controlled by $\sigma$.
- During E-step, find $q$ distribution:

$$\min_{q,c_{p,z}} KL(q||P_\theta) + \sigma \sum_{p,z} c_{p,z}$$

s.t. $E_q[\phi_{p,z}] \leq c_{p,z}$

where “c”s are maximums of expectation for each word tag pair by definition.
Primitive results

- Constrained model gives clustering that’s more sparse
- Clustering for a few phrases with 25 tags on BTEC ZH-EN

<table>
<thead>
<tr>
<th>Phrase/Word</th>
<th>Count of the most used tag</th>
<th>Number of tags used</th>
<th>Number of tags used</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>1194</td>
<td>1571</td>
<td>11</td>
</tr>
<tr>
<td>there is</td>
<td>53</td>
<td>50</td>
<td>5</td>
</tr>
<tr>
<td>‘d like</td>
<td>723</td>
<td>873</td>
<td>5</td>
</tr>
</tbody>
</table>
More experiments

- agreement constraint: different “good” models should agree on posterior distribution
- what model to agree with: another naïve Bayes model in the reverse direction or in the other language.
Agreement model

Implementation: multiply posteriors of two models together.
Agreement model

Only look at the first and last word of a phrase.

- implementation: multiply posteriors of two models together.
Agreement model

\[ \text{context} \rightarrow z \rightarrow \text{Phrase} \]

\[ \times \]

\[ \text{Phrase} \rightarrow z \rightarrow \text{context} \]

- Implementation: multiply posteriors of two models together.

Only look at the first and last word of a phrase.
Agreement model

- Implementation: multiply posteriors of two models together.
Outline

- Where do phrases come from?
- EM with posterior regularization
- results and future experiments
Evaluation through the translation pipeline on Urdu-English data
BLEU score, higher is better
Evaluation against supervised grammar
(Conditional Entropy, lower is better)
Confusion matrix against supervised labeling

EM

Agreement model between languages
Things we didn’t have time to get working

- Semi-supervised training with POS tags.
- Label single-word phrases with their POS tags.
Things we didn’t have time to get working

Bayesian Bayesian Bayesian Bayesian

• variational Bayes inference

Bayesian Bayesian Bayesian Bayesian

Bayesian Bayesian Bayesian
Things we didn’t have time to get working

- variational Bayes inference
Outline

- Where do phrases come from?
- EM with posterior regularization
- Results and future experiments

Thanks!
3:20pm Parametric models: posterior regularisation. Desai

3:35pm Training models with rich features spaces. Vlad

3:50pm Decoding with complex grammars. Adam

4:20pm Closing remarks. Phil

4:25pm Finish.
Discriminative Training

Vladimir Eidelman
Ziyuan Wang
Motivation

• Extract sparser features from grammars
  – Source Syntax
  – Target Syntax
  – Source Context
  – Glue Features
  – OOV
  – Backoff Rule
  – Morpheme construction

• ...
Source Syntactic Features

假期 快乐

Have a nice vacation.
到 现在 大约 六 个 小时 了 。

It's been about six hours now.
It's been about six hours now.
It's been about six hours now.
Glue Feature

S

X_5  X_{12}  X_{23}  \ldots  X_n
Glue Feature

Feature: $S_{X_5} = 1$
Glue Feature

Feature: $S \cdot X_{12} = 1$
Glue Feature

Feature: $S_{X_{23}} = 1$
Glue Feature

S

X_{12}

X_5

X_{17}

X_{11}

X_{25}

X_{23}

X_{25}
Glue Feature
Glue Feature

Feature: $\text{Glue}_S X_1 = 1$
Glue Feature

Feature: \( \text{Glue}_S X_{23} = 1 \)
Glue Feature

Feature: \( \text{Glue}_{S \times X_n} = 1 \)
Backoff

• In place of or combination with current dense feature
OOV

\[ X_1 \quad \text{Ragnarök} \]
\[ X_{23} \quad \text{supercalifragilisticexpialidocious} \]
\[ X_n \quad 6245 \]
OOV

$X_1$  
Ragnarök

$X_{23}$  
supercalifragilisticexpialidocious

$X_n$  
6245

Noun?  
Adjective?  
Number?
We want to...

- optimize model parameters to maximize translation quality on some metric (BLEU)
- do discriminative training so we can have features that directly help translation
- have thousands++ features
Motivation

• Minimum Error Rate Training
  – Does not scale well to more than handful of features
    • P(e) – Language Model
    • P(f|e) – Translation Rule
    • Pass through penalty

• Alternative approaches
  – Expected BLEU training
  – MIRA

• Evaluation
  – Language invariability (parameters, iterations, etc)
  – Standardizes comparison
# Training Comparison

<table>
<thead>
<tr>
<th></th>
<th>MERT</th>
<th>MIRA</th>
<th>Expected BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
<td>1-best</td>
<td>Margin-based</td>
<td>Probabilistic</td>
</tr>
<tr>
<td><strong>Objective</strong></td>
<td>Minimize error</td>
<td>Minimize loss augmented</td>
<td>Minimize expected error</td>
</tr>
<tr>
<td></td>
<td></td>
<td>score</td>
<td></td>
</tr>
<tr>
<td><strong>Optimization</strong></td>
<td>Line search</td>
<td>QP</td>
<td>Gradient based</td>
</tr>
<tr>
<td><strong>Limitations</strong></td>
<td>Direction of search</td>
<td>Approximation of reference</td>
<td>Approximate expectation</td>
</tr>
<tr>
<td></td>
<td>search unknown</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
女洗手间在哪里？

where's the ladies' room?
 Decomposable Objective

\( s(edge_i) = w f(edge_i) \)

where 's the ladies ' room?
女洗手间在哪里？

Decomposable Objective

\[ s(\text{edge}_i) = w \cdot f(\text{edge}_i) \]

where 's the ladies ' room?

\[ \sum_{\text{edge} \in \text{derivation}} s(\text{edge}) = S(e) \]
女洗手间在哪里？

Decomposable Objective

\( s(\text{edge}_i) = w f(\text{edge}_i) \)

Loss

approxBLEU

where 's the ladies ' room?

\[ \sum_{\text{edge} \in \text{derivation}} s(\text{edge}) = S(e) \]
女洗手间在哪里？

Decomposable Objective

\[ p(\text{edge}_i) = s(\text{edge}_i) = w \cdot f(\text{edge}_i) \]

Loss

\[ f(\text{edge}_i) = \text{approxBLEU} \]

Sentence level loss

Expected BLEU

\[ E_p[ f(\text{edge}_i) ] \]

MIRA

\[ f(\text{edge}_i) + p(\text{edge}_i) \]

where 's the ladies ' room ?

\[ \sum_{\text{edge} \in \text{derivation}} s(\text{edge}) = S(e) \]
Margin Infused Relaxed Algorithm (MIRA)

• Online Large-Margin Learning

• Crammer and Singer (2003)
  – Multi-class classification

• Taskar (2003)
  – Extension to structured value prediction

• Watanabe (2007), Chiang (2009)
  – Application to MT
Basic Learning Algorithm

Training data: $D = (x,y)$

$weight_0 = 0, total = 0, c = 0$

for iteration $1 \rightarrow n$

for $d = (x, y) \in D$

$weight_{c+1} = update \ weight_c \ with \ d$

$total = total + weight_{c+1}$

$c = c + 1$

$weight = \frac{total}{n \times size(D)}$
Update

\( s(x, y) = \mathbf{w} f(x, y) \)

- Learn \( \mathbf{w} \) so that correct outputs are given higher score than incorrect ones

\[
\min \| \mathbf{w}_{i+1} - \mathbf{w}_i \|
\]

- Keep the norm of the change to the weights as small as possible

- Subject to margin constraints:

\( s(x, y) - s(x, z) \geq Loss(y, z) \)

- Create margin between correct instance \( y \), and incorrect instance \( z \) at least as large as the Loss of \( z \)

- for all \( z \) which are possible labels of \( x \)
k-best MIRA

Training data: \( D = (e, f) \)

\[ \text{weight}_0 = 0, \text{total} = 0, c = 0 \]

for iteration 1 \( \rightarrow \) n

for \( d = (x, y) \in D \)

**Generate kbest( \( f \) ) = \{e\ldots e_k\}**

**Generate margin constraints \( \forall e \in \text{kbest}( f ) \)**

weight\( _{c+1} = \text{update weight}_c \text{ with } d \)

\( \text{total} = \text{total} + \text{weight}_{c+1} \)

\( c = c + 1 \)

\[ \text{weight} = \frac{\text{total}}{n \times \text{size}(D)} \]
女洗手间在哪里？

where's the ladies' room?
女洗手间在哪里？

where 's the ladies ' room? ||| LanguageModel=-6.3736... ||| 13.661
where 's the ladies? ||| LanguageModel=-5.76624... ||| 10.8657
where 's the ladies'? ||| LanguageModel=-6.51207 ... ||| 11.4501
where is the ladies' room? ||| LanguageModel=-7.18026 .. ||| 14.9181
where is the ladies? ||| LanguageModel=-6.5729... ||| 11.7432
where 's the ladies ' room ?  ||| LanguageModel=-6.3736....  ||| 13.661
where 's the ladies ?  ||| LanguageModel=-5.76624...  ||| 10.8657
where 's the ladies ' ?  ||| LanguageModel=-6.51207 ...  ||| 11.4501
where is the ladies ' room ?  ||| LanguageModel=-7.18026 ..  ||| 14.9181
where is the ladies ?  ||| LanguageModel=-6.5729...  ||| 11.7432

Model+BLEU

where is the ladies ' room ?  ||| LanguageModel=-7.18026  ||| 14.9181
where is the ladies ' ?  ||| LanguageModel=-7.31873||| 12.8778
where 's the ladies ' room ?  ||| LanguageModel=-6.3736||| 13.661
where is the ladies ?  ||| LanguageModel=-6.5729||| 11.7432
where 's the ladies ?  ||| LanguageModel=-5.76624  ||| 10.8657

where 's the ladies ' room ?
女洗手间在哪里？

where 's the ladies ' room? || LanguageModel=-6.3736... || 13.661
where 's the ladies? || LanguageModel=-5.76624... || 10.8657
where 's the ladies'? || LanguageModel=-6.51207... || 11.4501
where is the ladies ' room? || LanguageModel=-7.18026... || 14.9181
where is the ladies? || LanguageModel=-6.5729... || 11.7432

where is the ladies ' room? || LanguageModel=-7.18026 || 14.9181
where is the ladies'? || LanguageModel=-7.31873 || 12.8778
where 's the ladies ' room? || LanguageModel=-6.3736 || 13.661
where is the ladies? || LanguageModel=-6.5729 || 11.7432
where 's the ladies? || LanguageModel=-5.76624 || 10.8657

Model - BLEU

where is the bus depot for the ladies ' room? || LanguageModel=-10.7635 || 11.7463
where is the bus depot for the ladies? || LanguageModel=-10.1561 || 10.0082
where is the bus depot for the ladies'? || LanguageModel=-10.902... || 10.1763
where 's the ladies ' room? || LanguageModel=-6.3736 || 13.661
where is the bus depot for the ladies' room. || LanguageModel=-11.1228 || 10.8613
<table>
<thead>
<tr>
<th>Model</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>47.96</td>
</tr>
<tr>
<td>Model+BLEU</td>
<td>54.08</td>
</tr>
<tr>
<td>Model-BLEU</td>
<td>24.10</td>
</tr>
</tbody>
</table>
where 's the ladies ' room ? ||| LanguageModel=-6.3736.... ||| 13.661
where 's the ladies ? ||| LanguageModel=-5.76624... ||| 10.8657
where 's the ladies ' ? ||| LanguageModel=-6.51207 ... ||| 11.4501
where is the ladies ' room ? ||| LanguageModel=-7.18026 .. ||| 14.9181
where is the ladies ? ||| LanguageModel=-6.5729... ||| 11.7432

Oracle Translation

where is the ladies ' room ? ||| LanguageModel=-7.18026  ||| **14.9181**
where is the ladies ' ? ||| LanguageModel=-7.31873||| 12.8778
where 's the ladies ' room ? ||| LanguageModel=-6.3736||| 13.661
where is the ladies ? ||| LanguageModel=-6.5729||| 11.7432
where 's the ladies ? ||| LanguageModel=-5.76624 ||| 10.8657

where is the bus depot for the ladies ' room ? ||| LanguageModel=-10.7635 ||| 11.7463
where is the bus depot for the ladies ? ||| LanguageModel=-10.1561||| 10.0082
where is the bus depot for the ladies ' ? ||| LanguageModel=-10.902...||| 10.1763
where 's the ladies ' room ? ||| LanguageModel=-6.3736||| 13.661
where is the bus depot for the ladies ' room . ||| LanguageModel=-11.1228 ||| 10.8613
where 's the ladies ' room? ||| LanguageModel=-6.3736... ||| 13.661
where 's the ladies? ||| LanguageModel=-5.76624... ||| 10.8657
where 's the ladies'? ||| LanguageModel=-6.51207... ||| 11.4501
where is the ladies' room? ||| LanguageModel=-7.18026... ||| 14.9181
where is the ladies? ||| LanguageModel=-6.5729... ||| 11.7432

\[ \Delta f(e) \]

\[ Loss \]

where is the ladies' room? ||| LanguageModel=-7.18026... ||| 14.9181
where is the ladies'? ||| LanguageModel=-7.31873... ||| 12.8778
where 's the ladies' room? ||| LanguageModel=-6.3736... ||| 13.661
where is the ladies? ||| LanguageModel=-6.5729... ||| 11.7432
where 's the ladies? ||| LanguageModel=-5.76624... ||| 10.8657

where is the bus depot for the ladies' room? ||| LanguageModel=-10.7635... ||| 11.7463
where is the bus depot for the ladies? ||| LanguageModel=-10.1561... ||| 10.0082
where is the bus depot for the ladies'? ||| LanguageModel=-10.902... ||| 10.1763
where 's the ladies' room? ||| LanguageModel=-6.3736... ||| 13.661
where is the bus depot for the ladies' room. ||| LanguageModel=-11.1228... ||| 10.8613
\( \Delta f(e) \)

<table>
<thead>
<tr>
<th>Question</th>
<th>LanguageModel</th>
<th>Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where's the ladies' room?</td>
<td>-6.3736</td>
<td>13.661</td>
</tr>
<tr>
<td>Where's the ladies?</td>
<td>-5.76624</td>
<td>10.8657</td>
</tr>
<tr>
<td>Where's the ladies'?</td>
<td>-6.51207</td>
<td>11.4501</td>
</tr>
<tr>
<td>Where is the ladies' room?</td>
<td>-7.18026</td>
<td>14.9181</td>
</tr>
<tr>
<td>Where is the ladies?</td>
<td>-6.5729</td>
<td>11.7432</td>
</tr>
</tbody>
</table>

\( Loss \)

<table>
<thead>
<tr>
<th>Question</th>
<th>LanguageModel</th>
<th>Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where is the ladies' room?</td>
<td>-7.18026</td>
<td>14.9181</td>
</tr>
<tr>
<td>Where is the ladies'?</td>
<td>-7.31873</td>
<td>12.8778</td>
</tr>
<tr>
<td>Where's the ladies' room?</td>
<td>-6.3736</td>
<td>13.661</td>
</tr>
<tr>
<td>Where is the ladies?</td>
<td>-6.5729</td>
<td>11.7432</td>
</tr>
<tr>
<td>Where's the ladies?</td>
<td>-5.76624</td>
<td>10.8657</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Question</th>
<th>LanguageModel</th>
<th>Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Where is the bus depot for the ladies' room?</td>
<td>-10.7635</td>
<td>11.7463</td>
</tr>
<tr>
<td>Where is the bus depot for the ladies?</td>
<td>-10.1561</td>
<td>10.0082</td>
</tr>
<tr>
<td>Where is the bus depot for the ladies'?</td>
<td>-10.902</td>
<td>10.1763</td>
</tr>
<tr>
<td>Where's the ladies' room?</td>
<td>-6.3736</td>
<td>13.661</td>
</tr>
<tr>
<td>Where is the bus depot for the ladies' room.</td>
<td>-11.1228</td>
<td>10.8613</td>
</tr>
<tr>
<td>Question</td>
<td>LanguageModel</td>
<td>\Delta f(e)</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>---------------</td>
<td>---------------</td>
</tr>
<tr>
<td>Where is the ladies' room?</td>
<td><strong>-7.18026</strong></td>
<td>14.9181</td>
</tr>
<tr>
<td>Where is the ladies?</td>
<td><strong>-5.76624</strong></td>
<td>13.661</td>
</tr>
<tr>
<td>Where is the ladies' room?</td>
<td><strong>-6.3736</strong></td>
<td>13.661</td>
</tr>
<tr>
<td>Where is the ladies?</td>
<td><strong>-6.5729</strong></td>
<td>11.7432</td>
</tr>
<tr>
<td>Where is the bus depot for the ladies' room?</td>
<td><strong>-10.7635</strong></td>
<td>11.7463</td>
</tr>
<tr>
<td>Where is the bus depot for the ladies?</td>
<td><strong>-10.1561</strong></td>
<td>10.0082</td>
</tr>
<tr>
<td>Where is the bus depot for the ladies' room?</td>
<td><strong>-10.902...</strong></td>
<td>10.1763</td>
</tr>
<tr>
<td>Where is the ladies' room?</td>
<td><strong>-6.3736</strong></td>
<td>13.661</td>
</tr>
<tr>
<td>Where is the bus depot for the ladies' room?</td>
<td><strong>-11.1228</strong></td>
<td>10.8613</td>
</tr>
</tbody>
</table>
where 's the ladies ' room ? | | | LanguageModel=-6.3736... | | 13.661
where 's the ladies ? | | | LanguageModel=-5.76624... | | 10.8657
where 's the ladies ' ? | | | LanguageModel=-6.51207 ... | | 11.4501
where is the ladies ' room ? | | | LanguageModel=-7.18026 .. | | 14.9181
where is the ladies ? | | | LanguageModel=-6.5729... | | 11.7432

Δf(e)  
where is the ladies ' room ? | | | LanguageModel=-7.18026 | | 14.9181
where is the ladies ' ? | | | LanguageModel=-7.31873 | | 12.8778
where 's the ladies ' room ? | | | LanguageModel=-6.3736 | | 13.661
where is the ladies ? | | | LanguageModel=-6.5729 | | 11.7432
where 's the ladies ? | | | LanguageModel=-5.76624 | | 10.8657

Loss
where is the bus depot for the ladies ' room ? | | | LanguageModel=-10.7635 | | 11.7463
where is the bus depot for the ladies ? | | | LanguageModel=-10.1561 | | 10.0082
where is the bus depot for the ladies ' ? | | | LanguageModel=-10.902... | | 10.1763
where 's the ladies ' room ? | | | LanguageModel=-6.3736 | | 13.661
where is the bus depot for the ladies ' room . | | | LanguageModel=-11.1228 | | 10.8613
Online Updating

Learner 1

Learner 2

Learner 3

Sentence 23

Sentence 24

Sentence 25
Online Updating

Decode $w(i)$

Model

$+$BLEU

$-$BLEU

Update $w(i+1)$

Learner 1

Learner 2

Sentence 24

Learner 3

Sentence 25
Online Updating

Learner 1

Learner 2

Learner 3

Average

Final weights
Expected BLEU

• BLEU is just a geometric mean of ngram precisions
Expected BLEU

- Brevity penalty when hypothesis < reference
Expected BLEU

• Usually perform 1-best BLEU
  • argmax

• Expected BLEU replaces it argmax with sum
  • Function becomes continuous w.r.t weights

• Use approximate brevity penalty
  • Replace argmax with sum

• Differentiable
Expected BLEU

• Procedure:

1) LBFGS tuning for several iterations until convergence on the hypergraph
2) Re-decode the source data, generate updated hypergraph
3) Repeat
Preliminary Experiments

• Compare Expected BLEU with MIRA with equivalent grammar on same test set
• Incorporate fine-grained features
  – Source Syntax
  – Target Syntax
  – Source Context
  – Glue Features
  – OOV
  – Backoff Rule
Preliminary Results

![Bar chart showing preliminary results for MERT, ExpBleu, and ExpBleu+spare features. The bars are plotted against a y-axis representing scores ranging from 55 to 56.6. MERT has a score around 55.6, ExpBleu around 55.4, and ExpBleu+spare features has the highest score at around 56.4.](chart.png)
Coming Soon...

Decoding with Complex Grammars
3:20pm Parametric models: posterior regularisation. Desai

3:35pm Training models with rich features spaces. Vlad

3:50pm Decoding with complex grammars. Adam

4:20pm Closing remarks. Phil

4:25pm Finish.
Efficient Decoding for Synchronous Context-Free Grammars

Adam Lopez (Edinburgh)
Jonathan Graehl (ISI)
Chris Dyer (CMU)

with thanks to:
the whole workshop team,
Juri Ganitkevitch (JHU) & Jonny Weese (JHU)
Efficient Decoding for Synchronous Context-Free Grammars

Adam Lopez (Edinburgh)
Jonathan Graehl (ISI)
Chris Dyer (CMU)

with thanks to:
the whole workshop team,
Juri Ganitkevitch (JHU) & Jonny Weese (JHU)
The Story So Far
The Story So Far

- Induce a grammar.
The Story So Far

- Induce a grammar.
- Tune some model parameters.
The Story So Far

- Induce a grammar.
- Tune some model parameters.
- Get a BLEU score.
The Story So Far

- Induce a grammar.
- Tune some model parameters.
- *Decode a test set.*
- Get a BLEU score.
The Story So Far

- Induce a grammar.
- Tune some model parameters.
  - Decode a tuning set.
  - Decode a test set.
- Get a BLEU score.
The Story So Far

- Induce a grammar.
  - *Decode the training data.*
- Tune some model parameters.
  - *Decode a tuning set.*
- *Decode a test set.*
- Get a BLEU score.
The Price of Performance

- 1 Category (baseline): 20.8
- 25 Categories: 21.7
The Price of Performance

- 1 Category (baseline): 3.0 sec/sentence
- 25 Categories: 52 sec/sentence
Some Questions
Some Questions

- Why is it so slow?
Some Questions

• Why is it so slow?
• How can we speed it up?
Some Questions

- Why is it so slow?
- How can we speed it up?
- What’s the big idea?
Context-Free Grammar

\[ X^1 \rightarrow dianzi shang \]
\[ X^2 \rightarrow dianzi shang \]
\[ X^3 \rightarrow mao \]
\[ X^4 \rightarrow X^1 \text{ de } X^3 \]
\[ X^4 \rightarrow X^2 \text{ de } X^3 \]
\[ X^5 \rightarrow X^1 \text{ de } X^3 \]
\[ S \rightarrow X^4 \]
\[ S \rightarrow X^5 \]
Context-Free Grammar

\[ S \rightarrow X^4 \]
\[ S \rightarrow X^5 \]
\[ X^1 \rightarrow \text{dianzi shang} \]
\[ X^2 \rightarrow \text{dianzi shang} \]
\[ X^3 \rightarrow \text{mao} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3 \]
\[ X^4 \rightarrow X^2 \text{ de } X^3 \]
\[ X^5 \rightarrow X^1 \text{ de } X^3 \]
Context-Free Grammar

\[ X^1 \rightarrow \text{dianzi shang} \]
\[ X^2 \rightarrow \text{dianzi shang} \]
\[ X^3 \rightarrow \text{mao} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3 \]
\[ X^4 \rightarrow X^2 \text{ de } X^3 \]
\[ X^5 \rightarrow X^1 \text{ de } X^3 \]
\[ S \rightarrow X^4 \]
\[ S \rightarrow X^5 \]
Context-Free Grammar

$X^1 \rightarrow \text{dianzi shang}$
$X^2 \rightarrow \text{dianzi shang}$
$X^3 \rightarrow \text{mao}$
$X^4 \rightarrow X^1 \ \text{de} \ \ X^3$
$X^4 \rightarrow X^2 \ \text{de} \ \ X^3$
$X^5 \rightarrow X^1 \ \text{de} \ \ X^3$
$S \rightarrow X^4$
$S \rightarrow X^5$
Context-Free Grammar

\[
\begin{align*}
X^1 & \rightarrow dianzi shang \\
X^2 & \rightarrow dianzi shang \\
X^3 & \rightarrow mao \\
X^4 & \rightarrow X^1 \text{ de } X^3 \\
X^4 & \rightarrow X^2 \text{ de } X^3 \\
X^5 & \rightarrow X^1 \text{ de } X^3 \\
S & \rightarrow X^4 \\
S & \rightarrow X^5
\end{align*}
\]
Synchronous Context-Free Grammar

\[ X^1 \rightarrow \text{dianzi shang/the mat} \]
\[ X^2 \rightarrow \text{dianzi shang/mat} \]
\[ X^3 \rightarrow \text{mao/the cat} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3/X^3 \text{ on } X^1 \]
\[ X^4 \rightarrow X^2 \text{ de } X^3/X^3 \text{ of } X^2 \]
\[ X^5 \rightarrow X^1 \text{ de } X^3/X^1 \text{ 's } X^3 \]
\[ S \rightarrow X^4/X^4 \]
\[ S \rightarrow X^5/X^5 \]
Synchronous Context-Free Grammar

\[
\begin{align*}
X^1 & \rightarrow \text{dianzi shang/the mat} \\
X^3 & \rightarrow \text{mao/the cat} \\
X^4 & \rightarrow X^1 \text{ de } X^3/X^3 \text{ on } X^1 \\
S & \rightarrow X^5/X^5
\end{align*}
\]
Synchronous Context-Free Grammar

\[ X^1 \rightarrow \text{dianzi shang/the mat} \]
\[ X^3 \rightarrow \text{mao/the cat} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3/X^3 \text{ on } X^1 \]
\[ S \rightarrow X^5/X^5 \]

S \quad S
Synchronous Context-Free Grammar

\[ X^1 \rightarrow \text{dianzi shang/the mat} \]
\[ X^3 \rightarrow \text{mao/the cat} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3/X^3 \text{ on } X^1 \]
\[ S \rightarrow X^5/X^5 \]

S \------------------------------------------\ S
Synchronous Context-Free Grammar

\[ X^1 \rightarrow \text{dianzi shang/the mat} \]
\[ X^3 \rightarrow \text{mao/the cat} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3 / X^3 \text{ on } X^1 \]
\[ S \rightarrow X^5 / X^5 \]
Synchronous Context-Free Grammar

\[ X^1 \rightarrow \text{dianzi shang/the mat} \]
\[ X^3 \rightarrow \text{mao/the cat} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3/X^3 \text{ on } X^1 \]
\[ S \rightarrow X^5/X^5 \]

S \hline S
\hspace{3cm}
X^4 \hspace{3cm} X^4
Synchronous Context-Free Grammar

\[
\begin{align*}
X^1 & \rightarrow \text{dianzi shang/the mat} \\
X^3 & \rightarrow \text{mao/the cat} \\
X^4 & \rightarrow X^1 \text{ de } X^3/X^3 \text{ on } X^1 \\
S & \rightarrow X^5/X^5
\end{align*}
\]
Synchronous Context-Free Grammar

\[ X^1 \rightarrow \text{dianzi shang/the mat} \]
\[ X^3 \rightarrow \text{mao/the cat} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3/X^3 \text{ on } X^1 \]
\[ S \rightarrow X^5/X^5 \]
Synchronous Context-Free Grammar

\[ X^1 \rightarrow \text{dianzi shang/the mat} \]
\[ X^3 \rightarrow \text{mao/the cat} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3/X^3 \text{ on } X^1 \]
\[ S \rightarrow X^5/X^5 \]
Synchronous Context-Free Grammar

dianzi shang de mao
Synchronous Context-Free Grammar

\[ S \rightarrow X^4 \]
\[ X^4 \rightarrow X^1 \rightarrow \text{dianzi} \]
\[ X^4 \rightarrow X^3 \rightarrow \text{shang de mao} \]
Synchronous Context-Free Grammar

S

↓

X^4

↑

S

↓

X^4

X^1

X^3

dianzi shang de mao

dianzi shang de mao

the cat on the mat
Synchronous Context-Free Grammar

\[
\begin{align*}
S & \rightarrow X^4 \\
X^1 & \rightarrow \text{dianzi} \\
X^3 & \rightarrow \text{shang de mao}
\end{align*}
\]

\[
\begin{align*}
S & \rightarrow X^4 \\
X^3 & \rightarrow \text{the cat} \\
X^1 & \rightarrow \text{on the mat}
\end{align*}
\]
Parsing

$X^1 \rightarrow \text{dianzi shang/the mat}$
$X^2 \rightarrow \text{dianzi shang/mat}$
$X^3 \rightarrow \text{mao/the cat}$
$X^4 \rightarrow X^1 \text{ de } X^3/X^3 \text{ on } X^1$
$X^4 \rightarrow X^2 \text{ de } X^3/X^3 \text{ of } X^2$
$X^5 \rightarrow X^1 \text{ de } X^3/X^1 \text{ 's } X^3$
$S \rightarrow X^4/X^4$
$S \rightarrow X^5/X^5$

dianzi shang de mao
Parsing

\[ X^1 \rightarrow \text{dianzi shang} \]
\[ X^2 \rightarrow \text{dianzi shang} \]
\[ X^3 \rightarrow \text{mao} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3 \]
\[ X^4 \rightarrow X^2 \text{ de } X^3 \]
\[ X^5 \rightarrow X^1 \text{ de } X^3 \]
\[ S \rightarrow X^4 \]
\[ S \rightarrow X^5 \]

dianzi shang de mao
Parsing

\[ X^1 \rightarrow dianzi\ shang \]
\[ X^2 \rightarrow dianzi\ shang \]
\[ X^3 \rightarrow mao \]
\[ X^4 \rightarrow X^1\ de\ X^3 \]
\[ X^4 \rightarrow X^2\ de\ X^3 \]
\[ X^5 \rightarrow X^1\ de\ X^3 \]
\[ S \rightarrow X^4 \]
\[ S \rightarrow X^5 \]

\textit{dianzi\ shang\ de\ mao}
Parsing

$X^1 \rightarrow dianzi \ shang$
$X^2 \rightarrow dianzi \ shang$
$X^3 \rightarrow mao$
$X^4 \rightarrow X^1 \ de \ X^3$
$X^4 \rightarrow X^2 \ de \ X^3$
$X^5 \rightarrow X^1 \ de \ X^3$
$S \rightarrow X^4$
$S \rightarrow X^5$
Parsing

\[ X^1 \rightarrow \text{dianzi shang} \]
\[ X^2 \rightarrow \text{dianzi shang} \]
\[ X^3 \rightarrow \text{mao} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3 \]
\[ X^4 \rightarrow X^2 \text{ de } X^3 \]
\[ X^5 \rightarrow X^1 \text{ de } X^3 \]
\[ S \rightarrow X^4 \]
\[ S \rightarrow X^5 \]
Parsing

$X^1 \rightarrow dianzi \ shang$
$X^2 \rightarrow dianzi \ shang$
$X^3 \rightarrow mao$
$X^4 \rightarrow X^1 \ de \ X^3$
$X^4 \rightarrow X^2 \ de \ X^3$
$X^5 \rightarrow X^1 \ de \ X^3$
$S \rightarrow X^4$
$S \rightarrow X^5$

\[ \begin{array}{c}
0X_2^1 \\
\downarrow \\
dianzi \\
1 \\
\downarrow \\
shang \\
2 \\
\downarrow \\
de \ de \\
3 \\
\downarrow \\
mao \\
4
\end{array} \]
\[ X^1 \rightarrow \text{dianzi shang} \]
\[ X^2 \rightarrow \text{dianzi shang} \]
\[ X^3 \rightarrow \text{mao} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3 \]
\[ X^4 \rightarrow X^2 \text{ de } X^3 \]
\[ X^5 \rightarrow X^1 \text{ de } X^3 \]
\[ S \rightarrow X^4 \]
\[ S \rightarrow X^5 \]
 Parsing

\[
X^1 \to \text{dianzi shang}
\]
\[
X^2 \to \text{dianzi shang}
\]
\[
X^3 \to \text{mao}
\]
\[
X^4 \to X^1 \text{ de } X^3
\]
\[
X^4 \to X^2 \text{ de } X^3
\]
\[
X^5 \to X^1 \text{ de } X^3
\]
\[
S \to X^4
\]
\[
S \to X^5
\]
Parsing

$X^1 \rightarrow \text{dianzi shang}$

$X^2 \rightarrow \text{dianzi shang}$

$X^3 \rightarrow \text{mao}$

$X^4 \rightarrow X^1 \text{ de } X^3$

$X^4 \rightarrow X^2 \text{ de } X^3$

$X^5 \rightarrow X^1 \text{ de } X^3$

$S \rightarrow X^4$

$S \rightarrow X^5$
\[ X^1 \rightarrow \text{dianzi shang} \]
\[ X^2 \rightarrow \text{dianzi shang} \]
\[ X^3 \rightarrow \text{mao} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3 \]
\[ X^4 \rightarrow X^2 \text{ de } X^3 \]
\[ X^5 \rightarrow X^1 \text{ de } X^3 \]
\[ S \rightarrow X^4 \]
\[ S \rightarrow X^5 \]
\[ X^1 \rightarrow \text{dianzi shang} \]
\[ X^2 \rightarrow \text{dianzi shang} \]
\[ X^3 \rightarrow \text{mao} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3 \]
\[ X^4 \rightarrow X^2 \text{ de } X^3 \]
\[ X^5 \rightarrow X^1 \text{ de } X^3 \]
\[ S \rightarrow X^4 \]
\[ S \rightarrow X^5 \]
Parsing

S → X^5 → X^1, X^3 → dianzi, shang, de, mao

0S_4 → 0X_4^5 → 3X_4^3 → dianzi, shang, de, mao
dianzi  shang  de  mao

the  cat  on  the  mat
Analysis

Nodes: $O(Gn^2)$
Edges: $O(G^3n^3)$
Not so fast...

- Speed and memory footprint matter for both evaluation and tuning.
- What if $G$ is really big?
- What happens when we add an $n$-gram language model?
Not so fast...

• Speed and memory footprint matter for both evaluation and tuning.

• What if $G$ is really big?

• What happens when we add an $n$-gram language model?

$$\text{argmax}_{\text{English}} p(\text{Urdu}|\text{English})p(\text{English})$$
Language Models are Important
Language Models are Important

Impact on size of language model training data (in words) on quality of Arabic-English statistical machine translation system

AE BLEU[%]

LM trained on 219B words of web data
+LM Dynamic Programming
+LM Dynamic Programming

the ... mat
a ... mat
mat ... mat
+LM Dynamic Programming

the ... mat
a ... mat
mat ... mat

the ... cat
a ... cat
mat ... cat
+LM Dynamic Programming

the ... mat ———— the ... cat
a ... mat ———— a ... cat
mat ... mat ———— mat ... cat
+LM Dynamic Programming

the ... cat
a ... cat
mat ... cat
the ... mat
a ... mat
mat ... mat

the ... mat
a ... mat
mat ... mat
the ... cat
a ... cat
mat ... cat
+LM Dynamic Programming

Item (node) form: $X_{i,j,q,r}$
Cube Pruning Summary

- Parse Source
  - Result: -LM Hypergraph
- Incorporate n-grams bottom up, pruning +LM items along the way
  - Result: +LM Hypergraph
Experimental Sandbox

• Urdu 25 category grammar
• cdec decoder (Dyer et al., ACL 2010)
  • http://www.cdec-decoder.org
  • http://code.google.com/p/ws10smt
  • http://github.com/alopez/cdec
cdec

• Why it’s awesome:
  • Supports multiple models: linear chain CRF, SCFG, phrase-based
  • Generic hypergraph algorithms
  • Implements baseline: cube pruning
The Size of Source Forests

• 1 Category (baseline)
  • Edges per sentence: 188,954
  • Decode time per sentence: 3.0 seconds

• 25 Categories
  • Edges per sentence: 1,242,410
  • Decode time per sentence: 52 seconds
Oracle: Does -LM Pruning Help?

- Generate unpruned -LM graph
- Prune using inside-outside
- Cube pruning to obtain +LM graph
Oracle: Source Forest Pruning

global inside-outside pruned -LM forest (phrasal feature scores only)
Oracle: Tuned Source Forest Pruning

global inside-outside pruned -LM forest (phrasal feature scores only)
Coarse-to-fine Parsing

\[ X^1 \rightarrow \text{dianzi shang/the mat} \]
\[ X^2 \rightarrow \text{dianzi shang/mat} \]
\[ X^3 \rightarrow \text{mao/the cat} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3/X^3 \text{ on } X^1 \]
\[ X^4 \rightarrow X^2 \text{ de } X^3/X^3 \text{ of } X^2 \]
\[ X^5 \rightarrow X^1 \text{ de } X^3/X^1 's X^3 \]
\[ S \rightarrow X^4/X^4 \]
\[ S \rightarrow X^5/X^5 \]
Coarse-to-fine Parsing

$X^1 \rightarrow \text{dianzi shang/the mat}$
$X^2 \rightarrow \text{dianzi shang/mat}$
$X^3 \rightarrow \text{mao/the cat}$
$X^4 \rightarrow X^1 \text{ de } X^3/X^3 \text{ on } X^1$
$X^4 \rightarrow X^2 \text{ de } X^3/X^3 \text{ of } X^2$
$X^5 \rightarrow X^1 \text{ de } X^3/X^1 \text{ 's } X^3$
$S \rightarrow X^4/X^4$
$S \rightarrow X^5/X^5$

$X \rightarrow \text{dianzi shang/the mat}$
$X \rightarrow \text{dianzi shang/mat}$
$X \rightarrow \text{mao/the cat}$
$X \rightarrow X \text{ de } X/X \text{ on } X$
$X \rightarrow X \text{ de } X/X \text{ of } X$
$X \rightarrow X \text{ de } X/X \text{ 's } X$
$S \rightarrow X/X$
Coarse-to-fine Parsing

\[ X^1 \rightarrow \text{dianzi shang/the mat} \]
\[ X^2 \rightarrow \text{dianzi shang/mat} \]
\[ X^3 \rightarrow \text{mao/the cat} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3/X^3 \text{ on } X^1 \]
\[ X^4 \rightarrow X^2 \text{ de } X^3/X^3 \text{ of } X^2 \]
\[ X^5 \rightarrow X^1 \text{ de } X^3/X^1 \text{'s } X^3 \]
\[ S \rightarrow X^4/X^4 \]
\[ S \rightarrow X^5/X^5 \]
Coarse-to-fine Parsing

\[ X^1 \rightarrow \text{dianzi shang/the mat} \]
\[ X^2 \rightarrow \text{dianzi shang/mat} \]
\[ X^3 \rightarrow \text{mao/the cat} \]
\[ X^4 \rightarrow X^1 \text{ de } X^3/X^3 \text{ on } X^1 \]
\[ X^4 \rightarrow X^2 \text{ de } X^3/X^3 \text{ of } X^2 \]
\[ X^5 \rightarrow X^1 \text{ de } X^3/X^1 's X^3 \]
\[ S \rightarrow X^4/X^4 \]
\[ S \rightarrow X^5/X^5 \]

\[ X \rightarrow \text{dianzi shang/the mat} \]
\[ X \rightarrow \text{dianzi shang/mat} \]
\[ X \rightarrow \text{mao/the cat} \]
\[ X \rightarrow X \text{ de } X/X \text{ on } X \]
\[ X \rightarrow X \text{ de } X/X \text{ of } X \]
\[ X \rightarrow X \text{ de } X/X 's X \]
\[ S \rightarrow X/X \]

Good news: Grammar constant shrinks
Bad news: Can sometimes prune away all parses
Coarse-to-fine Parsing

BLEU

seconds/sentence

4.8 6.6 13.2 19.9 21.3 21.6 21.7 21.6
Coarse-to-fine Parsing

percent of forest retained

BLEU

0.03 0.06 0.15 0.37 0.48 0.64 0.96 1

4.8 6.6 13.2 19.9 21.3 21.6 21.7 21.6
Grammar Pruning

\[
X^1 \rightarrow \text{dianzi shang/the mat}
\]
\[
X^2 \rightarrow \text{dianzi shang/mat}
\]
\[
X^3 \rightarrow \text{mao/the cat}
\]
\[
X^4 \rightarrow X^1 \text{ de } X^3/X^3 \text{ on } X^1
\]
\[
X^4 \rightarrow X^2 \text{ de } X^3/X^3 \text{ of } X^2
\]
\[
X^5 \rightarrow X^1 \text{ de } X^3/X^1 \text{ 's } X^3
\]
\[
S \rightarrow X^4/X^4
\]
\[
S \rightarrow X^5/X^5
\]
Grammar Pruning

\[
\begin{align*}
X^1 & \rightarrow \text{dianzi shang/the mat} \\
X^2 & \rightarrow \text{dianzi shang/mat} \\
X^3 & \rightarrow \text{mao/the cat} \\
X^4 & \rightarrow X^1 \text{ de } X^3/X^3 \text{ on } X^1 \\
X^4 & \rightarrow X^2 \text{ de } X^3/X^3 \text{ of } X^2 \\
X^5 & \rightarrow X^1 \text{ de } X^3/X^3 \text{'s } X^3 \\
S & \rightarrow X^4/X^4 \\
S & \rightarrow X^5/X^5
\end{align*}
\]
Grammar pruning
(note: different data conditions)

BLEU

decoding time
Current Results Summary
Current Results Summary

- Source parse forest pruning works.
Current Results Summary

- Source parse forest pruning works.
- Can reduce overall decoding time by 40% with unpruned grammars.
Current Results Summary

- Source parse forest pruning works.
- Can reduce overall decoding time by 40% with unpruned grammars.
- Can reduce decoding time by an order of magnitude with pruned grammars (maybe at some cost in BLEU).
Current Results Summary

• Source parse forest pruning works.
• Can reduce overall decoding time by 40% with unpruned grammars.
• Can reduce decoding time by an order of magnitude with pruned grammars (maybe at some cost in BLEU).
• Ongoing work on more interesting algorithms...
Parse Forests as Grammars

Input: <dianzi shiang de mao, a cat on the mat>
Parse Forests as Grammars

Input: \(<\text{dianzi shiang de mao}, \text{a cat on the mat}>\)

**Isomorphic CFG**

\([X34] \rightarrow \text{a cat}\)
Parse Forests as Grammars

Input: <dianzi shiang de mao, a cat on the mat>

Isomorphic CFG

[X34] → a cat
[X02] → the mat
[X04a] → [X34] on [X02]
[X04a] → [X34] of [X02]
[X04b] → [X02]'s [X34]
[X04b] → [X02][X34]
[S] → [X04a]
[S] → [X04b]
Parse Forests as Grammars

Input: <dianzi shiang de mao, a cat on the mat>
Parse Forests as Grammars

Input: <dianzi shiang de mao , a cat on the mat>

Isomorphic CFG

[X34] → a cat
Translation is Intersection

- Translation by parsing is intersection
- Intersect source with grammar
- Yields a parse forest $\rightarrow$ target grammar
- Generate with target grammar
- Intersect with target language model
  (regular language)
Generation from Grammars
Isomorphic CFG

[X34] → *a cat*
[X02] → *the mat*

[X04a] → [X34] *on* [X02]
[X04a] → [X34] *of* [X02]
[X04b] → [X02] *’s* [X34]
[X04b] → [X02] [X34]

[S] → [X04a]
[S] → [X04b]
Generation from Grammars

**Isomorphic CFG**

\[ [X34] \rightarrow a \text{ cat} \]
\[ [X02] \rightarrow \text{ the mat} \]
\[ [X04a] \rightarrow [X34] \text{ on } [X02] \]
\[ [X04a] \rightarrow [X34] \text{ of } [X02] \]
\[ [X04b] \rightarrow [X02] \text{ 's } [X34] \]
\[ [X04b] \rightarrow [X02] [X34] \]
\[ [S] \rightarrow [X04a] \]
\[ [S] \rightarrow [X04b] \]

\[ S_{0,4}, \langle s \rangle, \langle s \rangle \rightarrow \bullet X_{0,4}^a \]
Generation from Grammars

Isomorphic CFG

\[ X^{a}_{0,4,\langle s \rangle,\langle s \rangle} \rightarrow \bullet X_{3,4} \text{ on } X_{0,2} \]

\[ S_{0,4,\langle s \rangle,\langle s \rangle} \rightarrow \bullet X^{a}_{0,4} \]
Generation from Grammars

Isomorphic CFG

\[
\begin{align*}
[X_{34}] & \rightarrow \text{a cat} \\
[X_{02}] & \rightarrow \text{the mat} \\
[X_{04a}] & \rightarrow [X_{34}] \text{ on } [X_{02}] \\
[X_{04a}] & \rightarrow [X_{34}] \text{ of } [X_{02}] \\
[X_{04b}] & \rightarrow [X_{02}] 's [X_{34}] \\
[X_{04b}] & \rightarrow [X_{02}] [X_{34}] \\
[S] & \rightarrow [X_{04a}] \\
[S] & \rightarrow [X_{04b}] \\
X_{0,4}',(s),(s) & \rightarrow \bullet X_{3,4} \text{ on } X_{0,2} \\
S_{0,4}',(s),(s) & \rightarrow \bullet X_{0,4}^a \\
X_{3,4}',(s),(s) & \rightarrow \bullet \text{a cat}
\end{align*}
\]
Generation from Grammars

**Isomorphic CFG**

\[
\begin{align*}
[X34] & \rightarrow a \text{ cat} \\
[X02] & \rightarrow the \text{ mat} \\
[X04a] & \rightarrow [X34] \text{ on} [X02] \\
[X04a] & \rightarrow [X34] \text{ of} [X02] \\
[X04b] & \rightarrow [X02] \text{'s} [X34] \\
[X04b] & \rightarrow [X02] [X34] \\
[S] & \rightarrow [X04a] \\
[S] & \rightarrow [X04b]
\end{align*}
\]

\[
\begin{align*}
X_{0,4}^a,\langle s\rangle,\langle s\rangle & \rightarrow \bullet X_{3,4} \text{ on } X_{0,2} \\
S_{0,4},\langle s\rangle,\langle s\rangle & \rightarrow \bullet X_{0,4}^a \\
X_{3,4},\langle s\rangle,\langle s\rangle & \rightarrow a \bullet \text{ cat} \\
X_{3,4},\langle s\rangle,\langle s\rangle & \rightarrow \bullet a \text{ cat}
\end{align*}
\]
Generation from Grammars

**Isomorphic CFG**

\[ [X02] \rightarrow \text{the mat} \]
\[ [X04a] \rightarrow [X34] \text{on} [X02] \]
\[ [X04a] \rightarrow [X34] \text{of} [X02] \]
\[ [X04b] \rightarrow [X02] \text{`}s [X34] \]
\[ [X04b] \rightarrow [X02] [X34] \]
\[ [S] \rightarrow [X04a] \]
\[ [S] \rightarrow [X04b] \]

\[ X_{0,4,\langle s\rangle,\langle s\rangle}^a \rightarrow \bullet X_{3,4} \text{ on } X_{0,2} \]

\[ S_{0,4,\langle s\rangle,\langle s\rangle} \rightarrow \bullet X_{0,4} \]

\[ X_{3,4,\langle s\rangle,\text{cat}} \rightarrow \text{a cat } \bullet \]
\[ X_{3,4,\langle s\rangle,a} \rightarrow \text{a } \bullet \text{ cat} \]
\[ X_{3,4,\langle s\rangle,\langle s\rangle} \rightarrow \bullet \text{a cat} \]
**Isomorphic CFG**

[X34] → *a cat*

[X02] → *the mat*

[X04a] → [X34] *on* [X02]

[X04a] → [X34] *of* [X02]

[X04b] → [X02] ‘*s* [X34]

[X04b] → [X02] [X34]

[S] → [X04a]

[S] → [X04b]

\[ X_{a}^{0,4,\langle s\rangle,\langle s\rangle} \rightarrow \bullet X_{3,4} \text{ on } X_{0,2} \]

\[ S_{0,4,\langle s\rangle,\langle s\rangle} \rightarrow \bullet X_{0,4}^{a} \]

\[ X_{3,4,\langle s\rangle,\text{cat}} \rightarrow \text{a cat } \bullet \]

\[ X_{3,4,\langle s\rangle,\text{a}} \rightarrow \text{a } \bullet \text{ cat} \]

\[ X_{3,4,\langle s\rangle,\langle s\rangle} \rightarrow \bullet \text{a cat} \]
Isomorphic CFG

\[ X_{3,4} \rightarrow a \text{ cat} \]
\[ X_{02} \rightarrow \text{the mat} \]
\[ X_{04a} \rightarrow [X_{34}] \text{ on } [X_{02}] \]
\[ X_{04a} \rightarrow [X_{34}] \text{ of } [X_{02}] \]
\[ X_{04b} \rightarrow [X_{02}] 's [X_{34}] \]
\[ X_{04b} \rightarrow [X_{02}] [X_{34}] \]
\[ S \rightarrow [X_{04a}] \]
\[ S \rightarrow [X_{04b}] \]

\[ S_{0,4}, \langle s \rangle, \langle s \rangle \rightarrow \bullet X_{0,4}^a \]
\[ X_{0,4}, \langle s \rangle, \langle s \rangle \rightarrow \bullet X_{3,4} \text{ on } X_{0,2} \]

\[ X_{3,4}, \langle s \rangle, \text{cat} \rightarrow a \text{ cat } \bullet \]
\[ X_{3,4}, \langle s \rangle, \text{a} \rightarrow a \bullet \text{ cat} \]
\[ X_{3,4}, \langle s \rangle, \langle s \rangle \rightarrow \bullet a \text{ cat} \]
Isomorphic CFG

\[
\begin{align*}
[X34] & \rightarrow a \ text{cat} \\
[X02] & \rightarrow the \ text{mat} \\
[X04a] & \rightarrow [X34] \ on \ [X02] \\
[X04a] & \rightarrow [X34] \ of \ [X02] \\
[X04b] & \rightarrow [X02] \ 's \ [X34] \\
[X04b] & \rightarrow [X02] \ [X34] \\
[S] & \rightarrow [X04a] \\
[S] & \rightarrow [X04b]
\end{align*}
\]
Isomorphic CFG

\[ [X_{34}] \rightarrow \text{a cat} \]
\[ [X_{02}] \rightarrow \text{the mat} \]
\[ [X_{04a}] \rightarrow [X_{34}] \text{ on } [X_{02}] \]
\[ [X_{04a}] \rightarrow [X_{34}] \text{ of } [X_{02}] \]
\[ [X_{04b}] \rightarrow [X_{02}] 's [X_{34}] \]
\[ [X_{04b}] \rightarrow [X_{02}] [X_{34}] \]
\[ [S] \rightarrow [X_{04a}] \]
\[ [S] \rightarrow [X_{04b}] \]

\[ X_{0,4}^a, \langle s \rangle, \text{cat} \rightarrow X_{3,4} \bullet \text{ on } X_{0,2} \]

\[ S_{0,4}, \langle s \rangle, \langle s \rangle \rightarrow \bullet X_{0,4}^a \]

\[ X_{3,4}, \langle s \rangle, \text{cat} \rightarrow \text{a cat} \bullet \]
\[ X_{3,4}, \langle s \rangle, \text{a} \rightarrow \text{a} \bullet \text{ cat} \]
\[ X_{3,4}, \langle s \rangle, \langle s \rangle \rightarrow \bullet \text{a cat} \]
Isomorphic CFG

\[ X_{0,4}, \langle s \rangle, \text{on} \rightarrow X_{3,4} \text{ on } \bullet X_{0,2} \]

\[ X_{0,4}, \langle s \rangle, \text{cat} \rightarrow X_{3,4} \bullet \text{ on } X_{0,2} \]

\[ X_{3,4}, \langle s \rangle, \text{cat} \rightarrow \text{a cat } \bullet \]

\[ X_{3,4}, \langle s \rangle, \text{a} \rightarrow \text{a } \bullet \text{ cat} \]

\[ X_{3,4}, \langle s \rangle, \langle s \rangle \rightarrow \bullet \text{a cat} \]
Isomorphic CFG

\[
X_{0,4,\langle s \rangle, \text{on}} \rightarrow X_{3,4} \text{ on } \bullet X_{0,2}
\]

\[
X_{0,2,\text{on,mat}} \rightarrow \text{the mat} \bullet
\]

\[
X_{0,4,\langle s \rangle, \text{cat}} \rightarrow X_{3,4} \bullet \text{ on } X_{0,2}
\]

\[
X_{3,4,\langle s \rangle, \text{cat}} \rightarrow \text{a cat} \bullet
\]

\[
X_{3,4,\langle s \rangle, \text{a}} \rightarrow \text{a } \bullet \text{ cat}
\]

\[
X_{3,4,\langle s \rangle, \langle s \rangle} \rightarrow \bullet \text{a cat}
\]
**Isomorphic CFG**

\[ \begin{align*} 
[X34] & \rightarrow \text{a cat} \\
[X02] & \rightarrow \text{the mat} \\
[X04a] & \rightarrow [X34] \text{ on } [X02] \\
[X04a] & \rightarrow [X34] \text{ of } [X02] \\
[X04b] & \rightarrow [X02] \text{'s } [X34] \\
[X04b] & \rightarrow [X02] [X34] \\
[S] & \rightarrow [X04a] \\
[S] & \rightarrow [X04b] \\
\end{align*} \]

**Generation from Grammars**

\[ \begin{align*} 
X_{0,4},\langle s\rangle,\text{mat} & \rightarrow X_{3,4} \text{ on } X_{0,2} \\
X_{0,4},\langle s\rangle,\text{on} & \rightarrow X_{3,4} \text{ on } \bullet X_{0,2} \\
X_{0,4},\langle s\rangle,\text{cat} & \rightarrow X_{3,4} \bullet \text{ on } X_{0,2} \\
S_{0,4},\langle s\rangle,\langle s\rangle & \rightarrow \bullet X_{0,4} \\
\end{align*} \]
Generation from Grammars

\[ X_{0,4,\langle s\rangle,\text{mat}} \rightarrow X_{3,4} \text{ on } X_{0,2} \]

\[ X_{0,2,\text{on,mat}} \rightarrow \text{the mat} \]

\[ X_{3,4,\langle s\rangle,\text{cat}} \rightarrow \text{a cat} \]
Generation from Grammars

Isomorphic CFG

\[ X_{0,4,\langle s\rangle,\text{mat}} \rightarrow X_{3,4} \text{ on } X_{0,2} \]

\[ X_{0,2,\text{on,mat}} \rightarrow \text{the mat} \]

\[ X_{3,4,\langle s\rangle,\text{cat}} \rightarrow \text{a cat} \]

- \([X34] \rightarrow \text{a cat}\)
- \([X02] \rightarrow \text{the mat}\)
- \([X04a] \rightarrow [X34] \text{ on } [X02]\)
- \([X04a] \rightarrow [X34] \text{ of } [X02]\)
- \([X04b] \rightarrow [X02] \text{ 's } [X34]\)
- \([X04b] \rightarrow [X02] [X34]\)
- \([S] \rightarrow [X04a]\)
- \([S] \rightarrow [X04b]\)
Theoretical Outcomes

- Works for arbitrary grammars, not just binary grammars
- Asymptotically faster than cube pruning ("hook trick", Liang et al. 2006).
- Produces lots of admissible heuristics (A*)
- No cube pruning: everything is monotonic.
Conclusions

- Faster algorithms are needed to make induced grammars practical.
- Workshop made significant progress towards this goal.
- More improvements are underway.
Outline

- 3:20pm Parametric models: posterior regularisation. Desai
- 3:35pm Training models with rich features spaces. Vlad
- 3:50pm Decoding with complex grammars. Adam
- 4:20pm Closing remarks. Phil
- 4:25pm Finish.
In this attack a large number of local residents has should vacate areas.

- Current state-of-the-art translation models struggle with language pairs which exhibit large differences in structure.
After this attack, a large number of local residents have to vacate the areas.

In this workshop we’ve made some small steps towards better translations for difficult language pairs.
After this attack, a large number of local residents have to vacate the areas.

In this workshop we’ve made some small steps towards better translations for difficult language pairs.

Google Translate:

*After the attack a number of local residents has blank areas.*
What we’ve achieved:

- The first unsupervised labelled SCFG induction algorithms:
  - by clustering translation phrases which occur in the same context we can learn which phrases are substituteable,
  - we have implemented parametric and non-parametric Bayesian clustering algorithms and shown positive results on real translation tasks.

- Improved SCFG decoders that efficiently decode grammars with many labels:
  - we have created faster search algorithms tuned for syntactic grammars.

- Discriminative training regimes to leverage features extracted from these grammars:
  - we’ve implemented two large scale discriminative algorithms for training our models.
What we’ve achieved:

- The first unsupervised labelled SCFG induction algorithms:
  - by clustering translation phrases which occur in the same context we can learn which phrases are substituteable,
  - we have implemented parametric and non-parametric Bayesian clustering algorithms and shown positive results on real translation tasks.

- Improved SCFG decoders that efficiently decode grammars with many labels:
  - we have created faster search algorithms tuned for syntactic grammars.

- Discriminative training regimes to leverage features extracted from these grammars:
  - we’ve implemented two large scale discriminative algorithms for training our models.

Thank you.