Unsupervised Multilingual Grammar Induction
• Languages exhibit variations in patterns of ambiguity
• Variations as natural supervision

בראשית בראש אלוהים את השמיים ואת הארץ
في البداية خلق الله السموات والأرض
בראשית ברא אלהים את השמיים ואת הארץ

في البداية خلق الله السموم والأرض

Morphology: acl 2008
POS tagging: emnlp 2008
naacl 2009
Syntax: acl 2009 (this talk)
Multilingual Cues

English: I saw the student from MIT
Multilingual Cues

English: I saw the student from MIT
Multilingual Cues

English: \( I \text{ saw } \text{the student from MIT} \)

Urdu: \( I \text{ MIT of student saw} \)
Multilingual Cues

English: [I saw] the student from MIT

Urdu: I MIT of student saw
Multilingual Cues

English: \([I \ saw] \text{the student from MIT}\)

Urdu: \(I \ MIT \ of \ student \ saw\)
Multilingual Cues

English:  

*I saw the student from MIT*

Urdu:  

*[I MIT]* of student saw?

Multilingual Cues

English: I saw the student from MIT

Urdu: [I MIT] of student saw
Multilingual Cues

English: I saw the student [from MIT]

Urdu: I MIT of student saw
Multilingual Cues

English:  \( I \) saw the student \( \text{[from MIT]} \)

Urdu:  \( I \) MIT of student saw
Multilingual Cues

English: \[ I \text{ saw [the student [from MIT]]} \]

Urdu: \[ I \text{ MIT of student saw} \]
Multilingual Cues

English:  

Urdu:  

I saw the student [from MIT]
Multilingual Cues

English: \[ I \text{ saw the student [from MIT ]} \]

Urdu: \[ I \text{ [[MIT of ] student] saw} \]
Multilingual Cues

English: \( I \) saw [the student [from MIT]]

Urdu: \( I \) [[MIT of] student] saw
Multilingual Cues

**English:**
\[ I \text{ saw } [\text{the student } [\text{from MIT }]] \]

**Urdu:**
\[ I [ [\text{MIT of } \text{student} ] \text{ saw} \]

**Main idea:** learn from systematic variations in phrase order and expression
Key Technical Challenge

Represent shared cross-lingual syntactic structure

- Linguistically plausible
  - Allow full range of syntactic divergence and translational freedom

- Computationally tractable
  - Support probabilistic operations: argmax, marginalization, sampling
Prior Representations

Synchronous Grammars [Wu 1997; Melamed 2003; Chiang 2005; Smith&Smith 2004; Eisner 2005; Blunsom et al 2008]

• Employed for modeling phrase reordering in MT
• In basic form, isomorphic trees (up to sibling order)

Node Matching [Burkett&Klein 2008]

• Ignores tree structure
• Marginalization is \#P-complete
Our Proposal

Probabilistic adaptation of *Unordered Tree Alignment* [Jiang et al 1995]

- Node alignments must respect tree structures
- Yet any number of nodes may remain unaligned
- Can marginalize and sample *all possible alignments* in linear time with dynamic program
For trees $T_1$ and $T_2$, an alignment $A$ is obtained in the following way:

1. Insert empty nodes into $T_1$ and $T_2$ and swap sibling order, until they are isomorphic

2. Overlay the resulting trees $T_1'$ and $T_2'$ to obtain $A$
For trees $T_1$ and $T_2$, an *alignment* $A$ is obtained in the following way:

1. Insert empty nodes into $T_1$ and $T_2$ and swap sibling order, until they are isomorphic

2. Overlay the resulting trees $T_1'$ and $T_2'$ to obtain $A$
For trees $T_1$ and $T_2$, an **alignment** $A$ is obtained in the following way:

1. Insert empty nodes into $T_1$ and $T_2$ and swap sibling order, until they are isomorphic

2. Overlay the resulting trees $T_1'$ and $T_2'$ to obtain $A$
John climbed Everest

English
John climbed Everest

English

John Everest on climbed

Urdu
John climbed Everest

English

John Everest on climbed

Urdu
A Generative Model

We observe:

```
DT  NN  VB  NNP
NN  NNP  VB
...  ...  ...
```
A Generative Model

We observe:

```
DT NN (VB NNP)
NN  NNP  VB
...
```
A Generative Model

We observe:

DT  NN  (VB  NNP)
NN  (NNP  VB)
...

...  ...  ...
A Generative Model

We observe:

```
DT  NN ( VB  NNP )
  NN  ( NNP  VB )
  ...  ...  ...
```
A Generative Model

We observe:

Hypothesize aligned trees that best explain:

- frequent POS sequence pairs
- lexical alignments
A Generative Model

We observe:

Hypothesize aligned trees that best explain:

- frequent POS sequence pairs
- lexical alignments

Parameters to learn
A Generative Model

We observe:

Hypothesize aligned trees that best explain:

- frequent POS sequence pairs
- lexical alignments

**Parameters to learn**

\[ \omega \]  Probability of constituent pairs of aligned nodes
A Generative Model

We observe:

Hypothesize aligned trees that best explain:

- frequent POS sequence pairs
- lexical alignments

Parameters to learn

\(\omega\) Probability of constituent pairs of aligned nodes
\(\phi^+\) Distribution on num. of word alignments between aligned nodes
\(\phi^-\) Distribution on num. of word alignments between unaligned nodes
A Generative Model

We observe:

Hypothesize aligned trees that best explain:

- frequent POS sequence pairs
- lexical alignments

Parameters to learn

- $\omega$: Probability of constituent pairs of aligned nodes
- $\phi^+$: Distribution on num. of word alignments between aligned nodes
- $\phi^-$: Distribution on num. of word alignments between unaligned nodes

(language-specific parameters for unaligned nodes [Klein&Manning 2002])
Generative Story

Draw alignment tree template \((T_1, T_2, A)\) from uniform distribution:

\[(\text{John climbed} \quad \text{Everest})\]

\[(\text{John Everest} \quad \text{on} \quad \text{climbed})\]
Generative Story

For each aligned node pair, draw a constituent pair jointly from $\omega$:
Generative Story

For each *aligned* node pair, draw a *constituent pair* jointly from $\omega$:

(left tree) (right tree)

(John climbed Everest) (John Everest on climbed)

NNP NNP
Generative Story

For each *aligned* node pair, draw a *constituent pair* jointly from ω:

(John climbed Everest) (John Everest on climbed)
Generative Story

For each aligned node pair, draw a constituent pair jointly from $\omega$:

(John climbed) (Everest)

(John Everest) on climbed

NNP NNP

VB NNP

NNP NNP IN VB
Generative Story

For each aligned node pair, draw a constituent pair jointly from $\omega$:
Generative Story

For each *unaligned* node, draw a *constituent* from language-specific parameters:

```
(John climbed Everest)
```

```
(John Everest on climbed)
```
Generative Story

For each *unaligned* node, draw a *constituent* from language-specific parameters:

(John climbed Everest)

(John Everest on climbed)
Generative Story

Draw word alignments between *aligned* and *unaligned* nodes according to $\phi^+$ and $\phi^-$:

(John climbed Everest)

(John Everest on climbed)
Generative Story

Draw word alignments between *aligned* and *unaligned* nodes according to $\phi^+$ and $\phi^-$: 

(John climbed Everest)

(John Everest on climbed)
Inference: Gibbs Sampling

• Sample each aligned tree pair conditioned on others:

\[ P \left( (T_1, T_2, A)_i \mid (T_1, T_2, A)_{\neg i} \right) \]

• Marginalize over all parameter values using standard closed forms (accumulated counts + hyperparameters)
Sampling Aligned Trees
Sampling Aligned Trees

- Hard to sample aligned tree pair: \((T_1, T_2, A)\)
Sampling Aligned Trees

- Hard to sample aligned tree pair: \((T_1, T_2, A)\)
- Use proposal distribution \(Q\), which assumes no nodes are aligned, to separately sample \(T_1^*, T_2^*\)
Sampling Aligned Trees

• Hard to sample aligned tree pair: \((T_1, T_2, A)\)

• Use proposal distribution \(Q\), which assumes no nodes are aligned, to separately sample \(T_1^*, T_2^*\)

• Accept with probability:

\[
\min \left\{ 1, \frac{P(T_1^*, T_2^*) Q(T_1, T_2)}{P(T_1, T_2) Q(T_1^*, T_2^*)} \right\} \text{ (Metropolis-Hastings)}
\]
Sampling Aligned Trees

• Hard to sample aligned tree pair: \((T_1, T_2, A)\)

• Use proposal distribution \(Q\), which assumes no nodes are aligned, to separately sample \(T_1^*, T_2^*\)

• Accept with probability:

\[
\min \left\{ 1, \frac{P(T_1^*, T_2^*)}{P(T_1, T_2)} \frac{Q(T_1, T_2)}{Q(T_1^*, T_2^*)} \right\} \text{ (Metropolis-Hastings)}
\]

• Conditionally sample tree alignment: \(A|T_1, T_2\)
Sampling Aligned Trees

• Hard to sample aligned tree pair: \((T_1, T_2, A)\)

• Use proposal distribution \(Q\), which assumes no nodes are aligned, to separately sample \(T_1^*, T_2^*\)

• Accept with probability:

\[
\min \left\{ 1, \frac{P(T_1^*, T_2^*) Q(T_1, T_2)}{P(T_1, T_2) Q(T_1^*, T_2^*)} \right\} \quad \text{(Metropolis-Hastings)}
\]

• Conditionally sample tree alignment: \(A | T_1, T_2\)
Sampling each Tree: Inside-Outside

• Recursively sample split-points from the top down

• Calculate probability of each split-point by marginalizing over all possible subtrees ("inside" table of inside-outside)

separately sample $T_1^*, T_2^*$

The boy ran through the haunted house
computing $P(T_1, T_2)$ \Rightarrow \text{need to marginalize over all possible alignments} \ A
computing $P(T_1, T_2)$ $\Rightarrow$ need to marginalize over all possible alignments $A$

- For $n_1 \in T_1, n_2 \in T_2$ table $D$ stores marginal probability of subtrees rooted at $n_1, n_2$

- Bottom-up dynamic program computes $D$ in time $O(|T_1||T_2|)$
computing $P(T_1, T_2)$ \implies$ need to marginalize over all possible alignments $A$

- For $n_1 \in T_1, n_2 \in T_2$ table $D$ stores marginal probability of subtrees rooted at $n_1, n_2$

- Bottom-up dynamic program computes $D$ in time $O(|T_1||T_2|)$

case 1:

![Diagram](image-url)
computing $P(T_1, T_2)$ \[\Rightarrow\] need to marginalize over all possible alignments $A$

- For $n_1 \in T_1, n_2 \in T_2$ table $D$ stores marginal probability of subtrees rooted at $n_1, n_2$

- Bottom-up dynamic program computes $D$ in time $O(|T_1||T_2|)$
computing $P(T_1, T_2)$ \iff need to marginalize over all possible alignments $A$

- For $n_1 \in T_1, n_2 \in T_2$ table $D$ stores marginal probability of subtrees rooted at $n_1, n_2$

- Bottom-up dynamic program computes $D$ in time $O(|T_1||T_2|)$

\text{case 3:}
computing $P(T_1, T_2)$ \implies$ need to marginalize over all possible alignments $A$

- For $n_1 \in T_1, n_2 \in T_2$ table $D$ stores marginal probability of subtrees rooted at $n_1, n_2$

- Bottom-up dynamic program computes $D$ in time $O(|T_1||T_2|)$

case 3:

\[ \lambda \]

similar for sampling $A|T_1, T_2$
Experiments

**Input:** Bilingual POS sequences (w/ giza alignments)

**Output:** Binary tree bracketings

**Evaluate:** Bracket precision, recall, F-measure, on held-out monolingual test data.

**Baseline:** (Bayesian) CCM [Klein & Manning 2002]
Corpora

- Korean-English Treebank: 5,000 sentences
- Urdu translation of WSJ: 4,300 sentences
  - no Urdu gold brackets
- English-Chinese Treebank: 3,850 sentences

Evaluate on various maximum sentence lengths (5 - 30)
KR with EN

F-measure

Max Sentence Length

Bilingual

Monolingual (CCM)
Results

• Average improvement across all scenarios:

  Precision:  +10
  Recall:     +8
  F-measure:  +9

• Average reduction in error relative to binary tree oracle: 19%
<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Percentage of Tree Nodes Aligned</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH-EN</td>
<td>71.6%</td>
</tr>
<tr>
<td>UR-EN</td>
<td>68.8%</td>
</tr>
<tr>
<td>KR-EN</td>
<td>60.2%</td>
</tr>
</tbody>
</table>
## Analysis

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Percentage of Tree Nodes Aligned</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH-EN</td>
<td>71.6%</td>
</tr>
<tr>
<td>UR-EN</td>
<td>68.8%</td>
</tr>
<tr>
<td>KR-EN</td>
<td>60.2%</td>
</tr>
</tbody>
</table>
Analysis

Percentage of tree nodes aligned

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH-EN</td>
<td>71.6%</td>
</tr>
<tr>
<td>UR-EN</td>
<td>68.8%</td>
</tr>
<tr>
<td>KR-EN</td>
<td>60.2%</td>
</tr>
</tbody>
</table>

Entropy of bracketed POS sequences

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>MONO</th>
<th>BI</th>
<th>GOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH (EN)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN (CH)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN (KR)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN (UR)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KR (EN)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analysis

Percentage of tree nodes aligned

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH-EN</td>
<td>71.6%</td>
</tr>
<tr>
<td>UR-EN</td>
<td>68.8%</td>
</tr>
<tr>
<td>KR-EN</td>
<td>60.2%</td>
</tr>
</tbody>
</table>

Entropy of bracketed POS sequences

- **MONO**
- **BI**
- **GOLD**

CH (EN) | EN (CH) | EN (KR) | EN (UR) | KR (EN)
Analysis

Percentage of tree nodes aligned

<table>
<thead>
<tr>
<th></th>
<th>CH-EN</th>
<th>UR-EN</th>
<th>KR-EN</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH-EN</td>
<td>71.6%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>UR-EN</td>
<td>68.8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KR-EN</td>
<td>60.2%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Entropy of bracketed POS sequences

<table>
<thead>
<tr>
<th></th>
<th>MONO</th>
<th>BI</th>
<th>GOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH (EN)</td>
<td>6.7</td>
<td>6.0</td>
<td>5.8</td>
</tr>
<tr>
<td>EN (CH)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN (KR)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EN (UR)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KR (EN)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The FCC effort Collapsed
The FCC effort Collapsed

Monolingual

X

DT NNP NN VBD
The FCC effort Collapsed

Monolingual  X

Bilingual (EN-UR)  ✓
The FCC effort Collapsed

Monolingual ✗

Bilingual (EN-UR) ✓

Pr_{mono} (NNP NN) < Pr_{bi} (NNP NN)
The FCC effort Collapsed

Monolingual X

Bilingual (EN-UR) ✓

\[ P_{\text{mono}}(\text{NNP NN}) < P_{\text{bi}}(\text{NNP NN}) \]

English: NNP NN

Urdu: NNP OF NN
Conclusions

**Key idea:** Use bilingual cues to learn better unsupervised monolingual models of grammar

- Adapt *Tree Alignment* to probabilistic setting:
  - Discover partial shared structure
  - Allow language-specific divergence
  - Computationally tractable

- Achieve improved performance on five corpora, across all sentence lengths
Thank you!
## Analysis

Entropy of constituent tag sequences

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>MONO</th>
<th>BI</th>
<th>GOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH-EN</td>
<td>6.6</td>
<td>5.6</td>
<td>5.3</td>
</tr>
<tr>
<td>CH-EN</td>
<td>6.9</td>
<td>5.9</td>
<td>5.5</td>
</tr>
<tr>
<td>UR-EN</td>
<td>6.2</td>
<td>6.2</td>
<td>6.9</td>
</tr>
<tr>
<td>KR-EN</td>
<td>6.8</td>
<td>5.9</td>
<td>5.6</td>
</tr>
<tr>
<td>UR-EN</td>
<td>6.8</td>
<td>6.2</td>
<td>5.9</td>
</tr>
<tr>
<td>avg</td>
<td>6.7</td>
<td>6.0</td>
<td>5.8</td>
</tr>
</tbody>
</table>
# Analysis

Entropy of constituent tag sequences

<table>
<thead>
<tr>
<th>Percentage of aligned tree nodes</th>
<th>MONO</th>
<th>BI</th>
<th>GOLD</th>
</tr>
</thead>
<tbody>
<tr>
<td>CH-EN</td>
<td>6.6</td>
<td>5.6</td>
<td>5.3</td>
</tr>
<tr>
<td>EN-CH</td>
<td>6.9</td>
<td>5.9</td>
<td>5.5</td>
</tr>
<tr>
<td>UR-EN</td>
<td>6.2</td>
<td>6.2</td>
<td>6.9</td>
</tr>
<tr>
<td>KR-EN</td>
<td>6.8</td>
<td>5.9</td>
<td>5.6</td>
</tr>
<tr>
<td>EN-KR</td>
<td>6.8</td>
<td>6.2</td>
<td>5.9</td>
</tr>
<tr>
<td>EN-UR</td>
<td>6.8</td>
<td>6.2</td>
<td>5.9</td>
</tr>
<tr>
<td><strong>avg</strong></td>
<td><strong>6.7</strong></td>
<td><strong>6.0</strong></td>
<td><strong>5.8</strong></td>
</tr>
</tbody>
</table>