Unsupervised Grammar Induction Using a Parent Based Constituent Context Model

Department of Computer Engineering
Sharif University of Technology, Tehran, Iran

S. A. Mirroshandel and G. Ghassem-Sani
mirroshandel@ce.sharif.edu sani@sharif.edu
Overview

- Grammar Induction
- Constituent Context Model
- Parent based Constituent Context Model
- Experimental Results
Grammar Induction
Grammar Induction
Grammar Induction

Hand-Craft Grammar

Grammar

NLP
A hand-crafted grammar is not usually completely satisfactory, and often fails to cover many unseen sentences.
Grammar Induction

Grammar

Grammar Induction

NLP

Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Availability of large corpora
Increasing the process power of computers
Success in other Statistical methods (POS tagging)
Grammar Induction

Supervised Grammar Induction

- Requires considerable amount of time and linguistic expertise
- Is more complicated for languages other than English (lack of resources)
- A hand-crafted treebank (and so the induced grammar) may be sensitive to a particular domain, application, or style
Grammar Induction
Grammar Induction

- Extraction of a language model out of a number of example sentences
- G.I. can be regarded as a learning method
  - It learns a grammar (usually probabilistic, e.g., a PCFG) from a training corpus
Why Is It So Hard?

NP (noun phrase) → ???

PLURAL NOUN

cats
Why Is It So Hard?

NP (noun phrase)  →  ???

DET  NOUN

a  cat

DET  NOUN

PLURAL NOUN
Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?

NP (noun phrase) → ???

Why Is It So Hard?
Why Is It So Hard?

NP (noun phrase) →

- DET
- ADJ
- NOUN
- PLURAL NOUN

- NP
- PP

Example: a cat in a sink
Why Is It So Hard?

NP (noun phrase) → ???

NP  CONJ  NP
  a cat   And   a dog

Diagram:

- DET
- ADJ
- NOUN
- PLURAL NOUN
- NP
- PP
- CONJ
Why Is It So Hard?

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
```

```
Why Is It So Hard?

Why Is It So Hard?
Two Steps in Grammar Induction

- Bracketing
- Labeling

- NP
- VP
- PP
Grammar Induction Categories

- Based on the type of corpora that different grammar induction methods use
  - Supervised
  - Unsupervised
  - Semi-Supervised

Grammar Induction Methods
Supervised Methods

- The training data set includes the Syntactic structure of the sample sentences, too.

\[ A_1 \rightarrow B_1, \quad A_2 \rightarrow B_2, \quad \ldots, \quad A_n \rightarrow B_n \]

Grammar
Unsupervised Methods

- The training data set includes only sentences with no more information.
- Usually tagged sentences are used for training.

\[
\begin{align*}
\text{unbracketed data set} & \quad \text{grammar} \\
\quad \quad A_1 & \rightarrow B_1 \\
\quad \quad A_2 & \rightarrow B_2 \\
\quad \quad \ldots & \\
\quad \quad A_n & \rightarrow B_n
\end{align*}
\]
Semi-Supervised Methods

- These methods use less supervision information than supervised ones.

- The training data set includes the Syntactic structure of some sentences or a part of structure of sentences.
Unsupervised Methods Applications

- In primary phases of constructing large treebanks
- In language modeling
- In some NLP research areas that do not require an exact grammar of sentences
Unsupervised Methods Categories

- Unsupervised Grammar Induction Methods
- Likelihood Based Methods
- Compression Based Methods
- Distribution Based Methods
Likelihood Based Methods

- This group of UGI selects maximum likelihood model using a PCFG
- Are also known as inside-outside
- Work using the EM algorithm
- Converge toward a local optimum state
- HIO and MIO are two examples of these methods that have shown satisfactory results, especially in Persian
Compression Based Methods

- These methods work using the minimum description length (MDL)
- The only factor with which these methods work is the compression of the most frequent sequence of tags
- Their results are not satisfactory
Distribution Based Methods

- These methods are based on a simple idea: “the sequences of words or tags that construct the same constituents appear in analogous contexts”

- Local context
  - CCM
  - CDC

- Global context
  - ABL
  - Emile
CCM Algorithm

- Was introduced by Klein and Manning
- Is a distribution based approach
- Works on the basis of: “constituents occur in their contexts”
- Distributed methods such as CCM do not work very well for free word order languages such as Persian
A sequences of tags is called “span”

For example in the following sentence, we have spans like NN1, NN1 NNS NNS VBD, VBD IN, VBD IN NN2
A span is called a “Constituent”, if it is included in a parse tree

For example in the following sentence, we have constituents such as NN1, NN1 NNS, VBD IN NN2, …
“Contexts” are pair of words that surround constituents.

E.g, $\diamond - NNS$ is a context for NN1.
Span, Constituent, and Context

```
Company incomes raised in February
```

- **Constituent**: Company, incomes, raised
- **Context**: in February
Span, Constituent, and Context

Company incomes raised in February

Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Span, Constituent, and Context

Company incomes raised in February

Constituent

Context
Span, Constituent, and Context

Constituent

Context

NP

S

VP

PP

NN1

NNS

VBD

IN

NN2

Company

incomes

raised

in

February
Span, Constituent, and Context

Company incomes raised

in February
Span, Constituent, and Context

Company incomes raised in February

- **Constituent**: Company, incomes, raised
- **Context**: in February

Diagram:
- S (Sentence)
  - VP (Verb Phrase)
    - PP (Prepositional Phrase)
      - IN (Preposition)
      - NN2 (Noun)
  - NP (Noun Phrase)
    - NN1 (Noun)
    - NNS (Noun)

Words:
- **NN**: Company, incomes, February
- **NNS**: raised
- **VBD**: raised
- **IN**: in
- **NN2**: raised

Analysis:
- The sentence is structured with a subject (Company) followed by a verb phrase (incomes raised) and a prepositional phrase (in February).
- The diagram illustrates the span, constituent, and context components of the sentence.
Span, Constituent, and Context

Company         incomes         raised         in       February
NN1            NNS           VBD            IN            NN2
Constituent
Context
Span, Constituent, and Context

Company         incomes         raised
            in       February

Constituent
Context
Bracketing & CCM

S

NP

NN1

VBD

IN

NN2

VP

0  Company  1  incomes  2  raised  3  in  4  February  5
Bracketing & CCM

Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Bracketing & CCM

Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Bracketing & CCM

Diagram:

- S
- VP
- PP
- NP
- NNS
- VBD
- IN
- NN1
- NN2

0 Company 1 incomes 2 raised 3 in 4 February 5

Grid:

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Bracketing & CCM
Bracketing & CCM

Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Bracketing & CCM

Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Bracketing & CCM

NP

NN1

VBD

IN

 NN2

VP

S

0 Company 1 incomes 2 raised 3 in 4 February 5
Bracketing & CCM
Bracketing & CCM
Bracketing & CCM

Non-Crossing Bracketing

Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Bracketing & CCM

Crossing Bracketing

Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Bracketing & CCM

Tree and Binary Tree Bracketing

[Diagram of a tree and binary tree with the words 'Company', 'incomes', 'raised', 'February' and 'in' indicated on the branches]
Constituent Context Model

- Two simple assumption:
  - Constituents of a parse do not cross each other
  - Constituents occur in constituent contexts
CCM Algorithm

- Only, binary tree bracketing is valid
- Uses EM algorithm to compute bracketing probabilities
  - In E step, according to some distribution \( P(B) \), a bracketing \( B \) is chosen, and then given that bracketing, the corresponding sentence is generated
  - In M step, algorithm tries to find a distribution for \( P(B) \) that maximize the probabilities
A Practical Example

- Consider the sentence

“Company incomes raised in February”
Different Trees

Company incomes raised in February
Different Trees

Company         incomes         raised
       in       February

Different Trees
Different Trees

Company         incomes         raised
in       February

NN1         NNS         VBD         IN         NN2
Different Trees

Company           incomes    raised      in      February

NN1               NNS        VBD         IN      NN2
Different Trees

Company incomes raised in February

Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Computing Probabilities

Diagram:

- **NN1**
- **NNS**
- **VBD**
- **IN**
- **NN2**

- Company
- incomes
- raised
- in
- February
Computing Probabilities

\[ P_{\text{constituency}}(\text{NN1 NNS}) \]
\[ P_{\text{context}}(\text{VBD}) \]
Computing Probabilities

\[ \text{Company} \quad \text{incomes} \quad \text{raised} \quad \text{in} \quad \text{February} \]
Computing Probabilities

\[ P_{\text{constituency}}(\text{IN NN2}) \]

\[ P_{\text{context}}(\text{VBD – ◊}) \]
Company incomes raised in February
Parent based Constituent Context Model

- Extending CCM algorithm, by adding some history notion of context and constituent information of each span's parent
Parent based CCM (PCCM)
Parent based CCM (PCCCM)
Parent based CCM (PCCM)

Company incomes raised in February
Parent based CCM (PCCM)

Constituent

Constituent’s Parent

Diagram:

S

VP

PP

NP

NN1

NNS

VBD

IN

NN2

Company

incomes

raised

in

February
Parent based CCM (PCCCM)

Company incomes raised in February

Constituent
Constituent’s Parent
Context
Parent based CCM (PCCCM)

Constituent
Constituent’s Parent
Context
Context’s Parent

Company         incomes         raised         in         February
Parent based CCM (PCCM)

Constituent

Constituent’s Parent

Context

Context’s Parent

Company         incomes         raised

in       February

NN

NNS

VBD

IN

NN2

Constituent’s Parent

Constituent

Context

Context’s Parent

S

VP

NP

PP

Constituent

Constituent’s Parent

Context

Context’s Parent

Company

incomes

raised

in

February

NN1

NNS

VBD

IN

NN2
Parent based CCM (PCCCM)
Parent based CCM (PCCCM)

Company incomes raised in February

- Constituent
- Constituent’s Parent
- Context
- Context’s Parent

S

VP

NP

VBD IN NN2

PP

Constituent’s Parent

Constituent

Context

Context’s Parent
Parent based CCM (PCCCM)
Parent based CCM (PCCM)

Constituent

Constituent’s Parent

Context

Context’s Parent

Company        incomes        raised        in        February

NN1            NNS            VBD            IN            NN2

Parent based CCM (PCCCM)

Constituent
Constituent’s Parent
Context
Context’s Parent

S
VP
PP

NN1  NNS  VBD  IN  NN2
Company  incomes  raised  in  February
Parent based CCM (PCCM)

Company         incomes         raised
in       February

Constituent
Constituent’s Parent
Context
Context’s Parent
Parent based CCM (PCCM)

Company
incomes
raised in February

Constituent
Constituent’s Parent
Context
Context’s Parent

Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Parent based CCM (PCCM)

Company incomes raised in February

Constituent
Constituent’s Parent
Context
Context’s Parent

Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Parent based CCM (PCCM)

Company incomes raised in February
Parent based CCM (PCCM)

Company         incomes         raised
in       February

Constituent
Constituent’s Parent
Context
Context’s Parent
Parent based CCM (PCCCM)
Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Parent based CCM (PCCCM)

Company         incomes         raised
in       February

Constituent
Constituent’s Parent
Context
Context’s Parent
Parent based CCM (PCCM)

Company       incomes       raised       in       February

Constituent
Constituent’s Parent
Context
Context’s Parent

Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Parent based CCM (PCCCM)

Constituent
Constituent’s Parent
Context
Context’s Parent

Company  incomes  raised  in  February

Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Parent based CCM (PCCCM)

Company         incomes         raised
                in       February

Constituent
Constituent’s Parent
Context
Context’s Parent

Constituent’s Parent
Context
Context’s Parent
Company incomes raised in February

Constituent's Parent

Context's Parent

Constituent
Parent based CCM (PCCM)

Constituent
Constituent’s Parent
Context
Context’s Parent

Company         incomes         raised         in         February
NN1               NNS                VBD                IN                NN2
Parent based CCM (PCCCM)

Constituent

Constituent’s Parent

Context

Context’s Parent

Company      incomes   raised   in    February
Context and Constituent in CCM vs. PCCM

Company incomes raised in February

NN1 NNS VBD IN NN2
Context and Constituent in CCM vs. PCCM

CCM

PCCM

Company incomes raised in February
Context and Constituent in CCM vs. PCCM

CCM

PCCM

Company incomes raised in February

Thursday 2008-07-24

Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Context and Constituent in CCM vs. PCCM

\[ P_{\text{constituency}}(NN1 \ NNS) \]
\[ P_{\text{context}}(\Diamond -VBD) \]
Context and Constituent in CCM vs. PCCM

$P_{\text{constituency}}(\text{NN1 NNS})$

$P_{\text{context}}(\text{VBD})$

CCM

PCCM

Company incomes raised in February
Context and Constituent in CCM vs. PCCM

CCM

\[ P_{\text{constituency}}(\text{NN1 NNS}) \]

\[ P_{\text{context}}(\text{VBD}) \]

PCCM

\[ P_{\text{constituency}}(\text{NN1 NNS} | \text{Parent(\text{NN1 NNS VBD IN NN2})}) \]

\[ P_{\text{context}}(\text{VBD} | \text{Parent(\text{VBD})}) \]
Context and Constituent in CCM vs. PCCM

CCM

PCCM

Company incomes raised in February
Context and Constituent in CCM vs. PCCM

CCM

PCCM

X

X

X

NN1

NNS

VBD

IN

NN2

Company

incomes

raised

in

February
Context and Constituent in CCM vs. PCCM

\[ P_{\text{constituency}}(IN \ NN2) \]

\[ P_{\text{context}}(VBD - \Diamond) \]
Context and Constituent in CCM vs. PCCM

CCM

\[ P_{\text{constituency}}(IN \ NN2) \]

\[ P_{\text{context}}(VBD - \Diamond) \]

PCCM

- Company
- incomes
- raised
- in
- February
Context and Constituent in CCM vs. PCCM

<table>
<thead>
<tr>
<th>CCM</th>
<th>PCCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_{\text{constituency}}(\text{IN} \ \text{NN2})$</td>
<td>$P_{\text{constituency}}(\text{IN} \ \text{NN2}</td>
</tr>
<tr>
<td>$P_{\text{context}}(\text{VBD} – \Diamond)$</td>
<td>$P_{\text{context}}(\text{VBD} – \Diamond</td>
</tr>
</tbody>
</table>

Thursday 2008-07-24
Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Time and Space Complexity

- Time Complexity of CCM is $O(n^3)$
  - Use of Dynamic Programming

- Space Complexity of CCM is $O(n^2)$

- Time Complexity of PCCM is $O(n^3)$
  - Use of memoization

- Space Complexity of PCCM is $O(n^3)$

- “$n$” is the number of words of the sentence
Experimental Results

- We ran CCM and PCCM on both English and Persian.

- Normally the structure of declarative sentences in Persian is "(S) (PP) (O) V". This language has high potential to be categorized in the free word order languages, especially in the preposition adjunction and complements.
Persian Free Word Orderness Example

- John dirooz be Mary ketâb dâd
  “John yesterday to Mary the book gave”
  “John gave the book to Mary yesterday”

- “John the book yesterday to Mary gave”

- “John to Mary yesterday the book gave”

- “yesterday John to Mary the book gave”

- “to Mary John yesterday the book gave”
Experimental Results

- English
Experimental Results

- English
- WSJ-10

Num. of sentences = 7422
Max. Len. = 10
Min. Len. = 2
Avg. Len. = 7
Num. of Words = 52248
Experimental Results

- English
  - WSJ-10
  - ATIS

**Air Travel Information System**

- Num. of sentences = 577
- Max. Len. = 35
- Min. Len. = 2
- Avg. Len. = 8
- Num. of Words = 4609
Experimental Results

- English
  - WSJ-10
  - ATIS
- Persian
Experimental Results

- English
  - WSJ
  - ATIS
- Persian
  - We constructed 3 corpuses from Peykareh

Num. of sentences = 3000
Max. Len. = 10
Min. Len. = 2
Avg. Len. = 7
Num. of Words = 22153
Experimental Results

- English
  - WSJ-10
  - ATIS
- Persian
  - We constructed 3 corpuses: Peykareh
    - First Corpus

Fixed Word Order: The sentences have been manually changed in such a way that the structure of "S PP O V" is held.
Experimental Results

- English
  - WSJ-10
  - ATIS

- Persian
  - We constructed two Persian corpuses from Peykareh
    - First Corpus
    - Second Corpus

*The sentences of the 2nd corpus have a high degree of “free-word-orderness”*
Experimental Results

- English
  - WSJ-10
  - ATIS

- Persian
  - We constructed
    - First Corpus
    - Second Corpus
    - Third Corpus

*Mixed Word Order: is produced by joining the two previous corpuses*
Experimental Results

- Using the ten-fold cross validation method

- PCCM has been Compared with
  - Original CCM: Klein and Manning
  - Context Distribution Clustering (CDC): Clark
  - EMILE and ABL: alignment based learning methods
  - LEFT and RIGHT Branching: based lines of the output
  - SUP-PCFG: Supervised PCFG parser with Viterbi parse

- The PARSEVAL Metrics (i.e., Precision, Recall, and F1-measure) were used for evaluation.
Experimental Results

- **Precision**

\[
UP(P, G) = \frac{\sum_i |brackets(P_i) \cap brackets(G_i)|}{\sum_i |brackets(P_i)|}
\]

- **Recall**

\[
UR(P, G) = \frac{\sum_i |brackets(P_i) \cap brackets(G_i)|}{\sum_i |brackets(G_i)|}
\]

- **F1**

\[
UF_1(P, G) = \frac{2}{UP(P, G)^{-1} + UR(P, G)^{-1}}
\]
Result on ATIS

<table>
<thead>
<tr>
<th>Model</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>EMILE</td>
<td>25.4</td>
</tr>
<tr>
<td>ABL</td>
<td>39.2</td>
</tr>
<tr>
<td>CDC-40</td>
<td>42</td>
</tr>
<tr>
<td>RBRANCH</td>
<td>42.9</td>
</tr>
<tr>
<td>CCM</td>
<td>51.2</td>
</tr>
<tr>
<td>PCCM</td>
<td>52.08</td>
</tr>
</tbody>
</table>
Result on WSJ-10

![Bar chart showing F1 scores for different models: LBRANCH, RANDOM, DEP-PCFG, RBRANCH, CCM, PCCM, SUP-PCFG, UBOUND. The scores range from 13 to 87.](chart.png)
Result on First Corpus (Persian)

![Graph showing F1 scores for different categories: LBRANCH, RBRANCH, CCM, PCCM, UBOUND.]

Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Result on Second Corpus (Persian)
Result on First Corpus (Persian)
Weaknesses of CCM

- Weakness in dealing with long sentences

- Weakness in dealing with free word order languages
Weaknesses of CCM (Cont.)

- CCM relies on repeated patterns of constituents and their contexts
- In both long sentences and free word order languages, the number of such patterns is highly decreased
- Consequently, there would be less information available during parsing
Why does PCCM work better?

- Considering parent information prevents from probability divergence, and thus parsing will be more-informative.

- But, when spans get longer, the co-occurrence of spans and their parents will substantially decrease!!
The effect of using parent information for different span's length

Persian First Corpus
Persian Second Corpus
Persian Third Corpus
English WSJ10

Thursday 2008-07-24
Unsupervised Grammar Induction Using a Parent Based Constituent Context Model
Smoothing

- Due to the usage of parent information, the smoothing task in PCCM is even more important than in CCM

- We used “additive smoothing” in PCCM
  - Adding 10 to each span and context
  - 2 as a constituent (valid context)
  - 8 as a distituent (invalid context)