The Hindi-Urdu Treebank

Lecture 7: 7/29/2011
Multi-representational, Multi-layered treebank

• Traditional approach:
  – Syntactic treebank: PS or DS, but not both
  – Layers are added one-by-one

• Our approach:
  – Syntactic treebank: both DS and PS
  – DS, PS, and PB are developed at the same time
  – Automatic conversion from DS+PB to PS

• Why?
  – DS and PS are both useful
  – Annotating them together allows us to maintain “consistency” and reduce annotation time
The team

- DS team: IIIT
- PB team: Univ of Colorado at Boulder
- PS team: UMass, Columbia Univ
- Conversion: Univ. of Washington

- Biweekly conference calls
- Group meetings every six months
Outline

• Overview of the treebank

• Three Representations
  – Dependency
  – Proposition Bank
  – Phrase Structure

• Conversion
Dependency structure (DS) and Phrase Structure (PS)

- **DS**: all nodes are labeled with words or empty strings

- **PS**: leaf nodes are labeled with words or empty strings, internal nodes are labeled with non-terminal symbols (special alphabet)
## Information in PS and DS

<table>
<thead>
<tr>
<th></th>
<th>PS (e.g., PTB)</th>
<th>DS (some target DS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS tag</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Function tag (e.g., -SBJ)</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Syntactic tag</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Empty category and co-indexation</td>
<td>Often yes</td>
<td>Often no</td>
</tr>
<tr>
<td>Allowing crossing</td>
<td>Often no</td>
<td>Often yes</td>
</tr>
</tbody>
</table>
Motivation 1: Two Representations

• Both phrase-structure treebanks and dependency treebanks are used in NLP
  – Collins/Charniak/Bikel parser for PS
  – CoNLL task on dependency parsing

• Problem: currently few treebanks (no?) with PS and DS which are independently motivated

➔ Our project: build treebank for Hindi/Urdu for which PS and DS are linguistically motivated from the outset
  – Dependency: Paninian grammar (Panini 400 BC)
  – Phrase structure: variant of Minimalism (Chomsky 1995)
Motivation 2:
Two Content Levels

• Everyone (?) wants syntax

• Recent popularity of PropBank (Palmer et al 2002): lexical predicate-argument structure; “semantics as surfacy as it gets”

• Recent experience: PropBank may inform some treebanking decisions

➡ Our project: build treebank with all levels from the outset
Goals

• Hindi/Urdu Treebank:
  – DS, PB, and PS for
    • 400K-word Hindi
    • 150K-word Urdu
  – Unified annotation guidelines
  – Frame files for PropBank

• Better understanding of DS=>PS conversion
Outline

• Overview of the project

• Three Representations
  – Dependency
  – Proposition Bank
  – Phrase Structure

• Conversion
There are 6 main *karakas (karaka relations)*:

- **karata (k1):** Activity of the verb resides in karta.
- **karma (k2):** Result of the verb resides in karma.
- **karana(k3):** Instrument helping in achieving the activity of the verb is karana
- **sampradaan (k4):** Receiver of the action is sampradaan
- **apaadan (k5):** Point of separation from which an entity has moved away in an action is apaadan
- **adhikaran (k7):** Place (k7p) or time (k7t) where the action is located
Full Set of Relations
Sample Paninian Analysis

raamane kala raata baazaara meM hari ko
'Ram' 'erg' 'yesterday' 'night' 'market' 'in' 'Hari' 'to'

titaaba dl
'book' 'gave'

de

k1 k7t k7p k4 k2

raama raata baazaara hari kitaaba

nmod

kala
## Basic Clause Structure

<table>
<thead>
<tr>
<th>अतिफ़</th>
<th>ने</th>
<th>किताब</th>
<th>को</th>
<th>पढ़ा</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atif</td>
<td>ne</td>
<td>kitaab</td>
<td>ko</td>
<td>paRhaa</td>
</tr>
<tr>
<td>Atif</td>
<td>Erg</td>
<td>book</td>
<td>Acc</td>
<td>read.Pfv</td>
</tr>
</tbody>
</table>

Atif read the book
Basic Clause Structure: DS

पढ़ा (read)

अतिफ़-ने (Atif)

किताब-को (book)
Outline

• Overview of the project

• Three Representations
  – Dependency
  – Proposition Bank
  – Phrase Structure

• Conversion
PropBank: Lexical Semantic Annotation

• Dependency annotation on top of DS
  – PropBank is a dependency representation, but the arc labels are different from DS

• Captures diathesis alternations:
  – John loaded the cart with hay.
  – John loaded hay on the cart.
  *hay* has same relation to predicate *load* in all these sentences

• PropBank annotates verb-meaning specific verbal roles
Basic Clause Structure: PropBank

पढ़ा Roleset: पढ़ना.01 (read)

अतिफ़-ने (Atif)

Arg0

किताब-को (book)

Arg1

पढ़ना.01

<table>
<thead>
<tr>
<th>Arg0</th>
<th>reader</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg1</td>
<td>what is read</td>
</tr>
</tbody>
</table>
Phrase Structure

• Inspired by Chomskyan Principles-and-Parameters approach

• (Mostly) binary branching

• Small number of non-terminals

• Key structural assumptions:
  – Only two marked argument positions for verbs, all other NPs are adjuncts and can appear anywhere
  – Use of traces for displacement from normal position
  – Case assigned under c-command
Basic Clause Structure: Phrase Structure

(Atif) read (book)
Unaccusatives

<table>
<thead>
<tr>
<th>दरवाज़ा</th>
<th>खुल</th>
<th>गया</th>
</tr>
</thead>
<tbody>
<tr>
<td>darwaaza</td>
<td>khul</td>
<td>gayaa</td>
</tr>
<tr>
<td>door</td>
<td>open</td>
<td>go.Pfv.MSg</td>
</tr>
</tbody>
</table>

The door opened.
Unaccusative: Dependency Structure

K1

दरवाज़ा (door)

खुल गया (open go)
Unaccusative: PropBank

खुल गया
(open go)

arg1

दरवाज़ा
(door)
Unaccusative: Phrase Structure

[Diagram showing a tree structure with nodes labeled as follows:
- VP
- NP
- VP-Pred
- *CASE*
- V
- (door)
- (open)
- (go)
Support Verb Constructions

<table>
<thead>
<tr>
<th>गहनें</th>
<th>चोरी</th>
<th>हो</th>
<th>गये</th>
</tr>
</thead>
<tbody>
<tr>
<td>geheneN</td>
<td>chorii</td>
<td>ho</td>
<td>gaye</td>
</tr>
<tr>
<td>jewels (m)</td>
<td>theft</td>
<td>do</td>
<td>go.Pfv.MPl</td>
</tr>
</tbody>
</table>

The jewels got stolen
Support Verb Constructions:
Dependency Structure

हो गये (do go)
k2  
गहनें (jewels)  
चोरी (theft)  
pof
Support Verb Constructions: PropBank

हो.sv (do)

<table>
<thead>
<tr>
<th>Arg0</th>
<th>agent of true predicate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arg1</td>
<td>true predicate</td>
</tr>
<tr>
<td>Arg2</td>
<td>patient of true predicate</td>
</tr>
</tbody>
</table>

हो Roleset: हो .sv

Arg2 (jewels)

Arg1 (theft)
Support Verb Constructions: Phrase Structure

(jewels)

(heft) (do go)
Where we are now

- Guidelines:
  - DS and PS guidelines are complete and checked
  - PropBank guidelines under development

- Annotation:
  - Finished 353K-word Hindi and 60k-word Urdu

- Automatic conversion from DS + PropBank in progress.

- Close co-operation in development of the three components essential