Identifying Urdu Complex Predication via Bigram Extraction

Miriam Butt\textsuperscript{1} Tina Bögel\textsuperscript{1} Annette Hautli\textsuperscript{1} Sebastian Sulger\textsuperscript{1} Tafseer Ahmed\textsuperscript{2}

\textsuperscript{1}University of Konstanz, Germany
\textsuperscript{2}University of Karachi, Pakistan

COLING 2012 in Mumbai, India
The situation

- Spoken and written language in Urdu/Hindi: heavy usage of complex predicates (CPs)
- Different types of CPs (Butt 1995):
  - Aspectual \( \text{V+V CPs: } \text{gir par-nə} \) ‘to fall suddenly (lit. fall fall)’
  - Permissive \( \text{V+V CPs: } \text{jane de-na} \) ‘to let go (lit. go give)’
  - \( \text{ADJ+V CPs: } \text{saf kar-nə} \) ‘to clean (lit. clean do)’
  - \( \text{N+V CPs: } \text{yad kar-nə} \) ‘to remember (lit. memory do)’
- In other languages:
  - \( \text{take a bite out of } X \) (lit. to bite X)
  - \( \text{give } X \text{ a stir} \) (lit. to stir X)
  - \( \text{außer Acht lassen} \) ‘to ignore (lit. let out of sight)’
- General problem in shallow and deep parsing approaches to Urdu/Hindi: proper treatment of complex predicates
The challenges

- Automatic distinction of CPs from simplex verbs
- Extraction of subcategorization frames
- Semantic role labeling
- Drawing semantic inferences
The challenges

- Automatic distinction of CPs from simplex verbs
- Extraction of subcategorization frames
- Semantic role labeling
- Drawing semantic inferences

Research questions:

- Can we blindly apply common statistical methods to extract the relevant patterns?
- Can we confirm existing theoretical hypotheses of N+V CP classes?
- Can visualization help us with this task?
Outline

1. Complex predicates
2. Methodology
3. Visualization
**N+V CPs**

- Combination of a *noun* which adds the main predicational content and a *light verb* which expresses subtle lexical semantic differences
- Highly productive constructions
- Proposal for different classes of N+V complex predicates based on a small case study (Ahmed and Butt 2011)
Combination of a *noun* which adds the main predicational content and a *light verb* which expresses subtle lexical semantic differences.

Highly productive constructions.

Proposal for different classes of \( \text{N+V} \) complex predicates based on a small case study (Ahmed and Butt 2011).

<table>
<thead>
<tr>
<th>N+V Type</th>
<th>Light Verb</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>class A</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>class B</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>class C</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
Class A: Psych predications (Noun + Light verb)

(1) \( \sqrt{ \text{Iar} \text{ki=ne kahani yad k-i} } \)
girl.F.Sg=Erg story.F.Sg.Nom memory.F.Sg.Nom do-Perf.F.Sg
‘The girl remembered a/the story.’
(lit.: ‘The girl did memory of the story.’)
N+V CPs

Class A: Psych predications (Noun + Light verb)

(2) \( \sqrt{\text{lar}ki=ne \ kahani \ yad \ k-i} \)
\( \text{girl.F.Sg=Erg story.F.Sg.Nom memory.F.Sg.Nom do-Perf.F.Sg} \)
‘The girl remembered a/the story.’
(lit.: ‘The girl did memory of the story.’)

\( \sqrt{\text{lar}ki=ko \ kahani \ yad \ he} \)
\( \text{girl.F.Sg=Dat story.F.Sg.Nom memory.F.Sg.Nom be.Pres.3P.Sg} \)
‘The girl remembers/knows a/the story.’
(lit.: ‘Memory of the story is at the girl.’)
N+V CPs

Class A: Psych predications (Noun + Light verb)

(3) ✓ larëki=ne kahani yad k-i
    girl.F.Sg=Erg story.F.Sg.Nom memory.F.Sg.Nom do-Perf.F.Sg
    ‘The girl remembered a/the story.’
    (lit.: ‘The girl did memory of the story.’)

✓ larëki=ko kahani yad he
    girl.F.Sg=Dat story.F.Sg.Nom memory.F.Sg.Nom be.Pres.3P.Sg
    ‘The girl remembers/knows a/the story.’
    (lit.: ‘Memory of the story is at the girl.’)

✓ larëki=ko kahani yad hu-i
    girl.F.Sg=Dat story.F.Sg.Nom memory.F.Sg.Nom become-F.Sg
    ‘The girl came to remember a/the story.’
    (lit.: ‘Memory of the story became to be at the girl.’)
**N+V CPs**

Class B: Agentive (transitive) CPs (Noun + Light verb)

(4) \(\sqrt[✓]{\text{bılal} = \text{ne makan tamir ki-ya}}\)
\ren\text{Bilal.M.Sg=Erg house.M.Sg.Nom construction.F.Sg do-Perf.M.Sg}\r
\r
‘Bilal built a/the house.’
Class B: Agentive (transitive) CPs (Noun + Light verb)

(5) \( \sqrt{\text{bılal=ne makan tamir ki-ya}} \)
Bilal.M.Sg=Erg house.M.Sg.Nom construction.F.Sg do-Perf.M.Sg
‘Bilal built a/the house.’

\[ \text{— *bılal=ko makan tamir hɛ} \]
Bilal.M.Sg=Dat house.M.Sg.Nom construction.F.Sg be.Pres.3.Sg


**Class B: Agentive (transitive) CPs (Noun + Light verb)**

\[(6) \quad \sqrt{bılal=ne \quad makan \quad tamir \quad ki-ya} \]

Bilal.M.Sg=Erg house.M.Sg.Nom construction.F.Sg do-Perf.M.Sg

‘Bilal built a/the house.’

— \[*bılal=ko \quad makan \quad tamir \quad hɛ\]

Bilal.M.Sg=Dat house.M.Sg.Nom construction.F.Sg be.Pres.3.Sg

— \[*bılal=ko \quad makan \quad tamir \quad hu-a\]

Bilal.M.Sg=Dat house.M.Sg.Nom construction.F.Sg become-M.Sg
Class c: Subject no undergoer (Noun + Light verb)

(7) √ bîlal=ne yîh fârṭ taslim ki
Bilal.M.Sg=Erg this condition.F.Sg acceptance.M.Sg do-Perf.F.Sg
‘Bilal accepted this condition.’
**N+V CPs**

**Class c: Subject no undergoer (Noun + Light verb)**

(8) √ bilal=ne yih fart taslim ki
Bilal.M.Sg=Erg this condition.F.Sg acceptance.M.Sg do-Perf.F.Sg
‘Bilal accepted this condition.’

√ bilal=ko yih fart taslim he
Bilal.M.Sg=Dat this condition.F.Sg acceptance.M.Sg be-3.Sg
‘Bilal accepted this condition.’
Class c: Subject no undergoer (Noun + Light verb)

(9) \[ \text{bilal}=\text{ne} \quad \text{yih} \quad \text{jasat} \quad \text{taslim} \quad \text{ki} \]
\[ \text{Bilal.M.Sg}=\text{Erg} \quad \text{this condition.F.Sg} \quad \text{acceptance.M.Sg} \quad \text{do-Perf.F.Sg} \]
‘Bilal accepted this condition.’

\[ \checkmark \quad \text{bilal}=\text{ko} \quad \text{yih} \quad \text{jasat} \quad \text{taslim} \quad \text{he} \]
\[ \text{Bilal.M.Sg}=\text{Dat} \quad \text{this condition.F.Sg} \quad \text{acceptance.M.Sg} \quad \text{be-3.Sg} \]
‘Bilal accepted this condition.’

???
\[ \text{bilal}=\text{ko} \quad \text{yih} \quad \text{jasat} \quad \text{taslim} \quad \text{hui} \]
\[ \text{Bilal.M.Sg}=\text{Dat} \quad \text{this condition.F.Sg} \quad \text{acceptance.M.Sg} \quad \text{become-F.Sg} \]
Our investigation

- Confirm the proposal by Ahmed and Butt (2011) with a larger empirical basis
- Extend the number of light verbs to four:
  1. *kar* ‘do’
  2. *ho* ‘be’
  3. *hū* ‘become’
  4. *rakh* ‘put’
- Start “naively” with commonly used statistical measures
- See whether these measures work for our data
Outline

1 Complex predicates
2 Methodology
3 Visualization
Extraction

Steps:
1. Use raw corpus of 7.9 million words harvested from the BBC Urdu website
2. Extract all bigrams which have one of the four light verbs as the right element
3. Data clean-up
4. Rank bigrams with the $X^2$ measure
5. Throw away bigrams with weak co-occurrence strength
6. Combine bigram lists to show the relative frequency of each noun with each light verb

<table>
<thead>
<tr>
<th>ID</th>
<th>Noun</th>
<th>Relative frequencies with light verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>kar</td>
</tr>
<tr>
<td>1</td>
<td>h2Asil ‘achievement’</td>
<td>0.771</td>
</tr>
<tr>
<td>2</td>
<td><em>a2</em>lAn ‘announcement’</td>
<td>0.982</td>
</tr>
<tr>
<td>3</td>
<td>bAt ‘talk’</td>
<td>0.853</td>
</tr>
<tr>
<td>4</td>
<td>SurUa2 ‘beginning’</td>
<td>0.530</td>
</tr>
</tbody>
</table>

Automatic transliteration as in Bögel (2012): unknown short vowels are represented as ‘*’
Hold-ups

- Spelling variation in Urdu words
- Inconsistent usage of “real” white space and zero-width non-joining
- Homonymy
  - $ki$ either feminine perfective form of $kar$ ‘do’ or genitive marker
- Homography
  - $kyA \rightarrow \text{‘that’}$, $kiyA \rightarrow \text{‘do.Perf.M.Sg’}$
- Nouns can be scrambled away from their light verbs
  - $\rightarrow$ Bigram approach helpless
- Light verbs can also be main verbs and auxiliaries in Urdu
  - $\rightarrow$ Much noise
Clustering

Automatic clustering of the data set

- Clusters based on the pattern of relative co-occurrence with the four light verbs
- **Problem:** How good are these clusters?
Automatic clustering of the data set

- Clusters based on the pattern of relative co-occurrence with the four light verbs
- Problem: How good are these clusters?

→ Visual analysis of the data set
Outline

1. Complex predicates
2. Methodology
3. Visualization
The concept

- Tight coupling of algorithms for automatic data analysis with visual components
- Eight visual variables: *position* (two variables x and y), *size*, *value*, *texture*, *color*, *orientation* and *shape*
- Exploit human perceptive abilities to support pattern detection

### Purpose of visualization

1. Overview of complex data sets
2. Starting point for an interactive exploration of data
3. Generation of new hypotheses, verification of existing hypotheses
Visualization – round 1

- Difficulty with detecting patterns among bare figures
- Requirement of a visual cue for the inspection of the clusters
- Difficulty of detecting patterns among bare figures
- Requirement of a visual cue for the inspection of the clusters
Visualization – round 1

- Mapping of relative frequencies to the visual variable *color*
- The higher the frequency, the darker the color
- Reference visualization of relative frequencies:

![Color mapping example](image)

- Proportional mapping between relative frequency and color
## Visualization – round 1

### Raw data

<table>
<thead>
<tr>
<th>Noun</th>
<th>kar</th>
<th>ho</th>
<th>hu</th>
<th>rakH</th>
</tr>
</thead>
<tbody>
<tr>
<td>h2Asil</td>
<td>0.771</td>
<td>0.222</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td><em>a2</em>lAn</td>
<td>0.982</td>
<td>0.011</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>bAt</td>
<td>0.853</td>
<td>0.147</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>SurUa2</td>
<td>0.530</td>
<td>0.384</td>
<td>0.086</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Visualization – round 1

Raw data

<table>
<thead>
<tr>
<th>Noun</th>
<th>kar</th>
<th>ho</th>
<th>hu</th>
<th>rakH</th>
</tr>
</thead>
<tbody>
<tr>
<td>h2Asil</td>
<td>0.771</td>
<td>0.222</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td><em>a2</em>lAn</td>
<td>0.982</td>
<td>0.011</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>bAt</td>
<td>0.853</td>
<td>0.147</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>SurUa2</td>
<td>0.530</td>
<td>0.384</td>
<td>0.086</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Visualized data
Visualization – round 1

Raw data

<table>
<thead>
<tr>
<th>Noun</th>
<th>kar</th>
<th>ho</th>
<th>hu</th>
<th>rakH</th>
</tr>
</thead>
<tbody>
<tr>
<td>h2Asil</td>
<td>0.771</td>
<td>0.222</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td><em>a2</em>IAn</td>
<td>0.982</td>
<td>0.011</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>bAt</td>
<td>0.853</td>
<td>0.147</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>SurUa2</td>
<td>0.530</td>
<td>0.384</td>
<td>0.086</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Visualized data

- Tool facilitates zooming and mousing over to see the underlying data set
Visualization – round 1

Raw data

<table>
<thead>
<tr>
<th>Noun</th>
<th>kar</th>
<th>ho</th>
<th>hu</th>
<th>rakH</th>
</tr>
</thead>
<tbody>
<tr>
<td>h2Asil</td>
<td>0.771</td>
<td>0.222</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td><em>a2</em>lAn</td>
<td>0.982</td>
<td>0.011</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>bAt</td>
<td>0.853</td>
<td>0.147</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>SurUa2</td>
<td>0.530</td>
<td>0.384</td>
<td>0.086</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Visualized data

- Tool facilitates zooming and mousing over to see the underlying data set
Visualization – round 1

Raw data

<table>
<thead>
<tr>
<th>Noun</th>
<th>kar</th>
<th>ho</th>
<th>hu</th>
<th>rakH</th>
</tr>
</thead>
<tbody>
<tr>
<td>h2Asil</td>
<td>0.771</td>
<td>0.222</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td><em>a2</em></td>
<td>0.982</td>
<td>0.011</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>bAt</td>
<td>0.853</td>
<td>0.147</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>SurUa2</td>
<td>0.530</td>
<td>0.384</td>
<td>0.086</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Visualized data

- Tool facilitates zooming and mousing over to see the underlying data set
Visualization – round 1

Raw data

<table>
<thead>
<tr>
<th>Noun</th>
<th>kar</th>
<th>ho</th>
<th>hu</th>
<th>rakH</th>
</tr>
</thead>
<tbody>
<tr>
<td>h2Asil</td>
<td>0.771</td>
<td>0.222</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td><em>a2</em>IAn</td>
<td>0.982</td>
<td>0.011</td>
<td>0.007</td>
<td>0.000</td>
</tr>
<tr>
<td>bAt</td>
<td>0.853</td>
<td>0.147</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>SurUa2</td>
<td>0.530</td>
<td>0.384</td>
<td>0.086</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Visualized data

- Tool facilitates zooming and mousing over to see the underlying data set
Visualization – round 1

Benefits of visualizing the initial clustering result

- At-a-glance detection of outliers, e.g. behavior of the verb $uTHA$ ‘to lift’

- Quick detection of clusters within clusters
- Visual evaluation of the goodness of the clustering
Visualization – round 1

**Result:**

- K-means clustering with $k=5$ best clustering algorithm according to the visualization
- Removal of clusters with consistently false hits (clusters 1, 3 and 4)
- Reduction of the list of bigrams from around 20,000 bigrams to 1,090
- Clusters 0 and 2 with many $N+V$ and $ADJ+V$ CPs are kept

**Next step:** Reclustering and visualization of the reduced data set
Visualization – round 2

Cluster 4:
- Much co-occurrence of item with $rak^h$ ‘put’
- Mixed cluster without complex predicates

Cluster 3:
- Items occur equally often with $kar$ ‘do’ and $ho$ ‘be’
- Cluster contains mostly ADJ+V sequences but hardly any CPs
Cluster 1:

- Occurs mostly with *ho* ‘be’ and *kar* ‘do’
- Cluster contains mostly *ADJ+V* sequences (also some valid *N+V* complex predicates)
- Interpreted as resultative constructions
Cluster 2:

- Largest cluster of all (around 600 members)
- Cluster 2 contains mostly $N+V$ sequences, but not all are $N+V$ CPs
- If $N+V$ CP, then of class B in Ahmed and Butt (2011) (no dative subjects allowed)
Cluster 0:

- Items occur mostly with *kar* ‘do’ and *ho* ‘be’
- Items also possible with *hu* ‘become’ (known from theoretical investigations)
- Contains valid $N+V$ complex predicates that correspond to Ahmed & Butt’s class A (psych predications)
### Result:

<table>
<thead>
<tr>
<th>N+V Type</th>
<th>Light Verb</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>kar ‘do’</td>
<td>ho ‘be’</td>
</tr>
<tr>
<td>CLASS A</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>CLASS B</td>
<td>+</td>
<td>–</td>
</tr>
<tr>
<td>CLASS C</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>

- N+V CPs of class A and B can be extracted from corpora
- Class C is difficult to detect
Discussion

Data sparsity

- Known $N+V$ combinations are not present in the corpus
- Problem of missing POS-tagged text for the language
Data sparsity

- Known \( N+V \) combinations are not present in the corpus
- Problem of missing POS-tagged text for the language

**BUT:**

- Partial confirmation of the \( N+V \) CP classes established by Ahmed and Butt (2011)
- Detection of \( A+V \) CPs
- Facilitation of data cleanup using visual keys
- Evaluation of clusters using methods from visualization
Future work

- Exploration of \( N + V \) and \( ADJ + V \) CPs in POS-tagged corpora (Urooj et al. 2012)
- Exploit existing information to extract scrambled \( N + V \) CPs
- Further extension of the visualization component:
  - Increasing the interaction with the data
  - Development of different methods for cluster visualization
Research question:

- Can we blindly apply common statistical methods to extract the relevant patterns?
Research question:

- Can we blindly apply common statistical methods to extract the relevant patterns?

  No, linguistic knowledge is required.
Summary

Research question:

- Can we blindly apply common statistical methods to extract the relevant patterns?
  
  No, linguistic knowledge is required.

- Can we confirm existing theoretical hypotheses of $N + V$ CP classes?
Research question:

- Can we blindly apply common statistical methods to extract the relevant patterns?
  
  No, linguistic knowledge is required.

- Can we confirm existing theoretical hypotheses of \( N + V \) CP classes?
  
  Yes, some clusters correspond to theoretically motivated CP classes.
Research question:

- Can we blindly apply common statistical methods to extract the relevant patterns?
  
  No, linguistic knowledge is required.

- Can we confirm existing theoretical hypotheses of $N + V$ CP classes?
  
  Yes, some clusters correspond to theoretically motivated CP classes.

- Can visualization help us with this task?
Research question:

- Can we blindly apply common statistical methods to extract the relevant patterns?
  No, linguistic knowledge is required.

- Can we confirm existing theoretical hypotheses of \( N+V \) CP classes?
  Yes, some clusters correspond to theoretically motivated CP classes.

- Can visualization help us with this task?
  Definitely!
Thank you!