Identifying Urdu Complex Predication via Bigram Extraction

Miriam Butt$^1$ Tina Bögel$^1$ Annette Hautli$^1$
Sebastian Sulger$^1$ Tafseer Ahmed$^2$

$^1$University of Konstanz, Germany
$^2$University of Karachi, Pakistan

COLING 2012 in Mumbai, India
The situation

- Spoken and written language in Urdu: heavy usage of complex predicates (CPs)
- Different types of CPs (Butt 1995):
  - aspectual V+V CPs: \textit{gir par} ‘suddenly fall’ (lit. fall fall)
  - permissive V+V CPs: \textit{jane di} ‘let go’ (lit. go give)
  - ADJ+V CPs: ‘ (lit. )’
  - N+V CPs: \textit{yad kar-na} ‘to remember (lit. memory do)’
- In other languages:
  - \textit{take a bite out of X} (lit. to bite X)
  - \textit{give X a stir} (lit. to stir X)
  - \textit{außer Acht lassen} ‘to ignore (lit. let out of sight)’
- General problem in shallow and deep parsing approaches to Urdu: 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)

<table>
<thead>
<tr>
<th>N+V Type</th>
<th>kar ‘do’</th>
<th>ḥe ‘be’</th>
<th>hu ‘become’</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLASS A</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>psych-predications</td>
</tr>
<tr>
<td>CLASS B</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>only agentive</td>
</tr>
<tr>
<td>CLASS C</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>subject is not an undergoer</td>
</tr>
</tbody>
</table>
Class A: Psych predications (Noun + Light verb)

(1) √  lahki=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.’)

√  lahki=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.’)

√  lahki=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.’)
Class B: Agentive (transitive) CPs (Noun + Light verb)

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

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

— \(\text{bılal=ko makan tamir hu-a} \)
\[\text{Bilal.M.Sg=Dat house.M.Sg.Nom construction.F.Sg become-M.Sg}\]
Class C: Subject no undergoer (Noun + Light verb)

(3)  

bilal=ne    yih  farț    taslim    ki
Bilal.M.Sg=Erg this condition.F.Sg acceptance.M.Sg do-Perf.F.Sg
‘Bilal accepted this condition.’

√  

bilal=ne    yih  farț    taslim  he
Bilal.M.Sg=Erg this condition.F.Sg acceptance.M.Sg be-3.Sg
‘Bilal accepted this condition.’

—  

bilal=ne    yih  farț    taslim  hui
Bilal.M.Sg=Erg this condition.F.Sg acceptance.M.Sg 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 on the right
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 (2162 instances)

<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><em>h2Asil</em> ‘achievement’</td>
<td>0.771</td>
</tr>
<tr>
<td>2</td>
<td><em>a2</em>!An ‘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><em>SurUa2</em> ‘beginning’</td>
<td>0.530</td>
</tr>
</tbody>
</table>

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

- Nouns can be scrambled away from their light verbs
  → Bigram approach helpless
- Light verbs can also be main verbs and auxiliaries in Urdu
  → Lots of noise to the left of them
- Homonymy
  - *ki* either feminine perfective form of *kar-na* ‘to do’ or genitive marker
- Spelling variation in Urdu words
- Inconsistent usage of “real” white space and zero-width non-joiner
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
- Exploit human perceptive abilities to support pattern detection
- Eight visual variables: position (two variables x and y), size, value, texture, color, orientation and shape

Purpose of visualization

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

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

- 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:

  - Proportional mapping between relative frequencies 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>

Visualized data

![Visualized data table]

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>IA2n</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>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

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 mostly \( N+V \) and \( \text{ADJ}+V \) CPs are kept
→ Reclustering and visualization of the reduced data set
Visualization – round 2

Cluster 3:
- Occurs equally with *kar* ‘do’ and *ho* ‘be’
- Cluster contains mostly **ADJ** + **V** CPs

Cluster 4:
- Mixed cluster without **N** + **V** complex predicates
Result:

- Again five clusters with k-means (verified by cluster visualization)
- Cluster 4 items occur mainly with $rak^h$ and contains many false hits
- Cluster 3 kar ‘do’ and ho ‘be’ and
Visualization – round 2

Cluster 0:
- Occurs mostly with *kar* ‘do’ and *ho* ‘be’
- Contains valid N+V complex predicates

Cluster 1:
- Occurs mostly with *ho* ‘be’ and *kar* ‘do’
- Cluster contains valid N+V complex predicates as well as ADJ+V CPs
Visualization – round 2

- Cluster 2 contains mostly nouns, some of which are $N + V$ CPs.
- $N + V$ CPs of class B in Ahmed and Butt (2011) (no dative subjects allowed)
Discussion

Data sparsity

- Known $N+V$ combinations are not present in the corpus
Discussion

- Problem of missing POS-tagged text for the language

**BUT:**

- Partial confirmation of the $N+V$ CP classes established by Ahmed and Butt (2011)
- Successful evaluation of clusters using methods from visualization
- Facilitation of data cleanup using visual keys
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.
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?
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?
  Yes, some clusters correspond to theoretically motivated CP classes.
- Can visualization help us with this task?
  Definitely!
Thank you!