Automatic Extraction of Complex Predicates in Bengali

Dipankar Das, Santanu Pal, Tapabrata Mondal, Tanmoy Chakraborty and Sivaji Bandyopadhyay

Department of Computer Science & Engineering
Jadavpur University, India

Multiword Expressions: from Theory to Applications 2010
Bengali Complex Predicates (CPs)

- **Compound Verb (CompV)**
  - [verb] + verb
  - *mere phela* ‘kill’

- **Conjunct Verb (ConjV)**
  - [noun/ adjective] + verb
  - *bharsha kara* ‘to depend’, *jhakjhak kara* ‘to glow’
Compound Verb (CompV)

- Full Verb (FV) + Light Verb (LV)

  - *Full Verb (FV)*
    - Conjunctive participial form –e/iya
    - Infinitive form –te/ite

  - *Light Verb (LV)*
    - Bears inflection based on *Tense, Aspect and Person*
    - *Polysemous, Semantically bleached*
    - Confined into some definite (e.g. 15) candidate seeds (Paul, 2010)
Conjunct Verb (ConjV)

- Noun/Adjective + Light Verb (LV)
- Uninflected noun/adjective
- Light Verb (LV)
  - Bears inflection based on Tense, Aspect and Person
Complex Predicates (CPs) are MWE

- Absence of conventional meaning of the *Light Verbs* (Sinha, 2009)
  
  *mere phela* [kill-drop] ‘to kill’ (*CompV*)  
  *bharsha kara* [depend-do] ‘to depend’ (*ConjV*)

- Typical Case
  
  - *dekha kara* [see-do] ‘to meet’
  - Follows similar lexical pattern “*Full Verb + Light Verb*”
  - Both *Full Verb* and *Light Verb* loose their conventional meanings
  - Generate a completely different meaning (‘to meet’)

---

28/08/2010

MWE 2010
Constraints

Butt (1993) and Paul (2004)

- **Control Construction (CC)**
  - *likhte bollo* ‘asked to write’
  - *likhte badhyo korlo* ‘forced to write’

- **Modal Control Construction (MCC)**
  - *jete hobe* ‘have to go’
  - *khete hobe* ‘have to eat’
Constraints

- **Passives (Pass)**
  - *dhora porlo* ‘was caught’
  - *mara holo* ‘was beaten’

- **Auxiliary Construction (AC)**
  - *bose ache* ‘is sitting’
  - *niye chilo* ‘had taken’
Constraints

Mukherjee et al. (2006)

- **Serial Verb (SV)**
  - *niye gelo* ‘take-go’ (took and went)
  - *diye gelo* ‘give-go’ (gave and went)
  - Similar lexical patterns \( FV + LV \)
  - Not mono-clausal
  - Both *Full Verb* (FV) and *Light Verb* (LV) behave like independent syntactic entities
Applications

- WordNet (Miller et al., 1990)
- VerbNet (Kipper-Schuler, 2005)*
- Parsing
- Machine Translation
Earlier Attempts

- *South Asian languages* (Abbi, 1991; Bashir, 1993; Verma, 1993)
- Kashmiri (Kaul, 1985)
- Oriya (Mohanty, 1992)
- Urdu (Butt, 1995)
- Hindi (Burton-Page, 1957; Hook, 1974; Mukherjee *et al.*, 2006; Chakrabarti *et al.*, 2008; Sinha, 2009)
- Bengali (Sarkar, 1975; Paul, 2004; Paul, 2009; Paul, 2010)
Bengali

- No automatic method in Bengali
- Morphologically rich
- Less computerized

- Sixth popular language in the World
  (http://www.ethnologue.com/ethno_docs/distribution.asp?by=size)

- Second in India
- National language of Bangladesh
System’s Outline

○ Identification of Complex Predicates (CPs)
○ Identification of Lexical Scope
○ Evaluation
Identification of Complex Predicates (CPs)

- Preparation of Corpora
- Extracting Complex Predicates (CPs)
Preparation of Corpora (1/2)

- Rabindranath Rachanabali
  (www.rabindra-rachanabali.nltr.org/)

- EILMT travel and tourism corpus
  “Development of English to Indian languages Machine Translation”

  The EILMT project is funded by the Department of Information Technology (DIT), Ministry of Communications and Information Technology (MCIT), Government of India

- 400 Development Sentences
- 500 Test Sentences
Preparation of Corpora (2/2)

- Open source Bengali shallow parser
  (http://ltrc.iiit.ac.in/showfile.php?filename=downloads/shallow_parser.php)

- Shallow parser gives different morphological information
  - root, lexical category of the root, gender, number, person, case, vibhakti, tam, suffixes etc.
Identification of Complex Predicates (CPs)

- Preparation of Corpora
- Extracting Complex Predicates (CPs)
Extracting Complex Predicates (1/6)

Lexical Pattern

- **Compound Verb**
  - \{[XXX](v) [YYY] (v)\}

- **Conjunct Verb**
  - \{[XXX] (n/adj) [YYY] (v)\}
Extracting Complex Predicates
(2/6)

Compound Verb

- Identify root forms of the *Light Verbs (LVs)* from shallow parsed result
- Another table stores root forms and corresponding dictionary forms of *Light Verbs*
- The table contains total 378 verb entries
- The dictionary forms of *Light Verbs (LVs)* are retrieved from Table
Compound Verb

- Check for Light Verb \((LV)\) seed list (Paul, 2004)

- All are in dictionary forms
Extracting Complex Predicates (4/6)

Compound Verb

- Check for *Full Verb* (*FV*)
  Conjunctive participial form -e/iya
  or
  Infinitive form –te/ite

- Shallow morphological information (e.g. suffixes of the verb)
Extracting Complex Predicates (5/6)

Compound Verb

- If ((dictionary forms of *Light Verb* (LVs) is present in the list of *Light Verbs*) && (*Full Verb* (FVs) contains suffixes of *e/iya* || –te/ite))

→ Compound verbs (*CompVs*)
Conjunct Verb

- Check for **Light Verb** \((LV)\) from frequently used seed list

```
deoya ‘give’  kara ‘do’
nëoya ‘take’  luga ‘start’
paoya ‘pay’  kata ‘cut’
```

- If \(\text{basic POS of the root of first words ([XXX])} = \text{“noun” } \text{|| “adj”} \) \&\& \(\text{basic POS of the following word ([YYY]) is “verb”}\) → Conjunct Verb \((ConjV)\)
System’s Outline

- Identification of Complex Predicates
- Identification of Lexical Scope
- Evaluation
Identification of Lexical Scopes (1/6)

Ex1:

\[ \text{chalte giye 'walking' + pore gelam 'fell down'} \]

\[ \text{giye pore = 'reached'} \]

I <fell down while walking>.
Identification of Lexical Scopes (2/6)

Ex1:

আমি  চলতে  গিয়ে  পারে  গলাম
(ami) (chalte) (giye) (pore) (gelam).

I <fell down while walking>.

[chalte]=[(chal)+(-te)] ‘walk’

আমি  চলতে  গিয়ে  পারে  গলাম
(ami) (chalte) (giye) (pore) (gelam).

I <fell down while walking>.

[chalte giye]=[(chal)+(-te)]+[giye (jaoya ‘go’)] ‘walking’
Identification of Lexical Scopes (3/6)

Ex1:

আমি চলতে গিয়ে গরে গেলাম
(ami) (chalte) (giye) (pore) (gelam).
I <fell down while walking>.

[pore] = [(por) + (-e)] ‘fell’

আমি চলতে গিয়ে গরে গেলাম
(ami) (chalte) (giye) (pore) (gelam).
I <fell down while walking>.

[pore gelam] = [(por) + (-e)] + [gelam (jaoya ‘go’)] ‘fell down’
Identification of Lexical Scopes
(4/6)

Ex1:

[chalte giye ‘walking’ + pore gelam ‘fell down’]

Starting point of Search:

- Why from the beginning of a Verb Group?
- Why not from the first identified Light Verb?
Identification of Lexical Scopes (5/6)

Ex2:

utable dekhlam ‘get up and saw’

utable pore dekhlam ‘get up and saw’

utable pore ‘get up’ or pore dekhlam ‘fall and see’
Identification of Lexical Scopes (6/6)

\[\text{[uteh pore 'get up'] = [(uth)+(-e)]+[pora 'fall']}\]

\[\text{dekhram 'saw'}\]
System’s Outline

- Identification of Complex Predicates (CPs)
- Identification of Lexical Scope
- Evaluation
Evaluation (1/5)

- Manual
- Two different corpus
- Total 800 development and 1000 test sentences
- Error Analysis using Constraint Validation
- Identification of lexical scopes increases the number of Complex Predicates (CPs)
Evaluation (2/5)

- Diagnostic test
- Complex Predicates vs. Constraints for Non-Complex Predicates (Butt, 1993; Paul, 2004)
- Confusion Matrix (in %)

|       | CompV | ConjV | NMCV | CC   | MCC  | Pass | AC  |
|-------|-------|-------|------|------|------|------|-----|
| CompV | 0.76  | 0.00  | 0.02 | 0.00 | 0.00 | 0.03 | 0.02|
| ConjV | 0.04  | 0.72  | 0.03 | 0.01 | 0.02 | 0.02 | 0.02|
| NMCV  | 0.17  |       | 0.65 | 0.00 | 0.02 | 0.02 | 0.02|
| CC    | 0.01  | 0.00  | 0.00 | 0.56 | 0.01 | 0.02 | 0.02|
| MCC   | 0.00  | 0.00  | 0.00 | 0.07 | 0.65 | 0.00 | 0.02|
| Pass  | 0.12  | 0.01  | 0.00 | 0.00 | 0.00 | 0.78 | 0.00|
| AC    | 0.06  | 0.07  | 0.04 | 0.00 | 0.00 | 0.08 | 0.54|
Evaluation (3/5)

- Non-MonoClausal Verbs (NMCV), Passives (Pass) and Auxiliary Construction (AC) are identified as compound verbs (CompVs)
- Non-MonoClausal Verbs (NMCV) and Auxiliary Construction (AC) occur as conjunct verbs (ConjVs)
- Reason: Frequencies of non-Complex Predicates (non-CPs) are reasonably higher in the corpora

Frequencies of Instances in two corpora
Evaluation (4/5)

Recall, Precision and F-Scores on two corpora before and after lexical scope determination (test set)

| EILMT          | Recall | Precision | F-Score |
|----------------|--------|-----------|---------|
| Compound Verb (CompV) | 65.92% | 80.11%    | 72.32%  |
|                | 70.31% | 82.06%    | 75.73%  |
| Conjunct Verb (ConjV) | 94.65% | 80.44%    | 86.96%  |
|                | 96.96% | 83.82%    | 89.90%  |

| Rabindra Rachanabali          | Recall | Precision | F-Score |
|-------------------------------|--------|-----------|---------|
| Compound Verb (CompV)         | 68.75% | 81.81%    | 74.71%  |
|                               | 72.22% | 84.61%    | 77.92%  |
| Conjunct Verb (ConjV)         | 94.11% | 83.92%    | 88.72%  |
|                               | 95.23% | 84.71%    | 89.66%  |
Evaluation (5/5)

- System suffers from clausal detection
- Passives (Pass) and Auxiliary Construction (AC) requires the knowledge of argument structure
- Problem in Conjunct Verbs
Conclusion

- Identification of Bengali Complex Predicates
- Proposed automatic methods for extraction
- Lexical Scope Identification
- Diagnostic tests for evaluation

- Only lexical patterns are insufficient
- Clause Identification
- Special focus on conjunct verbs
- Adoption of subcategorization frames or argument structures for solving the error issues
Thank you 😊