Classification of Verbs – Towards Developing Bengali Verb Subcategorization Lexicon

Somnath Banerjee, Dipankar Das and Sivaji Bandyopadhyay
Department of Computer Science & Engineering
Jadavpur University, Kolkata-700032, India
The NLP Research Group at Jadavpur University

- Computer Science and Engineering Department
- Teaching
  - NLP in both Undergraduate and Postgraduate courses in Computer Science and Engineering
- Research
NLP Research

- National Consortium R & D Projects
  - Cross Lingual Information Access
  - English to Indian Languages Machine Translation Systems
  - Indian Languages to Indian Languages Machine Translation Systems
NLP Research

- International Research Projects
  - “Advanced Platform for Question Answering Systems” with Prof. Patrick Saint Dizier, France
  - “Sentiment Analysis” with Prof. Junichii Tsujii and Prof. Manabu Okumura, Japan
  - “Answer Validation through Textual Entailment” with Prof. Alexander Gelbukh, Mexico
NLP Research Areas

- Answer Validation Through Textual Entailment
- Machine Translation using EBMT approach (Manipuri – English)
- Opinion Mining and Opinion Summarization
- Emotion Analysis
- Temporal Relations Extraction
NLP Research Areas

- Named Entity Extraction
- Subcategorization Frame Acquisition
- Question Answering Systems
- Text Summarization
- Statistical Machine Translation
- Language Generation
- Multi Word Extraction
Outline

- Motivation
- Previous work
- Language Challenges
- Bengali Subcategorization Frames Acquisition
- Classification of Bengali Verbs
- Evaluation
- Future Task
Motivation

- Subcategorization refers to certain kinds of relations between words and phrases in a sentence.
- A subcategorization frame is a statement of what types of syntactic arguments a verb (or adjective) takes, such as objects, infinitives, that-clauses, participial clauses and subcategorized prepositional phrases.
Motivation

- No existing parser in Bengali
  - Subcategorization frame information helps in parsing
- Subcategorization frames
  - Phrase alignment in a SMT system
  - Question answering systems
- To build verb subcategorization lexicon for Bengali using English VerbNet and Bengali-English Bilingual Dictionary
Outline

- Motivation
- Previous work
- Language Challenges
- Bengali Subcategorization Frames Acquisition
- Classification of Bengali Verbs
- Evaluation
- Future Task
Previous work (1/5)

○ ANLT (Boguraev and Briscoe, 1987)
  - Alvey Natural Language Tools
  - Manually prepared machine readable subcategorization lexicon

○ ACQUILEX (Copestake, 1992)
  - Acquisition of Lexical Knowledge for NLP Systems
  - Designed to support representation of multilingual lexical information extracted from machine readable dictionaries

○ COMLEX Syntax (Grishman et al., 1994)
  - Computational lexicon consisting of syntactical information for approximately 38,000 English headwords
Previous work (2/5)

- **FrameNet (Baker et al., 1998)**
  - On-line lexical resource for English, based on frame semantics and supported by corpus evidence.
  - Range of semantic and syntactic combinatory possibilities (valences) of each word in each of its senses
  - More than 11,600 lexical units, more than 6,800 of which are fully annotated in more than 960 semantic frames exemplified in more than 150,000 annotated sentences.

- **PropBank (Palmer et al., 2005)**
  - A corpus annotated with verbal propositions and their arguments
  - Wide popularity for the semantic role labeling and NLP tasks.
  - PropBank differs from FrameNet
    - It commits to annotating all verbs in its data.
    - All arguments to a verb must be syntactic constituents.
Previous Work (3/5)

- **Valex (Korhonen *et al.*, 2006)**
  - 163 Subcategorization Frame (SCF) types, superset of ANLT and COMLEX
    - Provides lexical entry for each verb and SCF combination
    - Total 212,741 entries, 33 per verb on average
    - Suitable for statistical NLP, linguistic and psycholinguistic use
Previous work (4/5)

- **VerbNet** (Kipper-Schuler, 2006)
  - VerbNet associates the semantics of a verb with its syntactic frames
  - Hierarchical domain-independent, broad-coverage verb lexicon based on Levin (1993) verb classes
  - Mappings to other lexical resources such as WordNet (Miller, 1990; Fellbaum, 1998), XTAG (XTAG Research Group, 2001), and FrameNet (Baker et al., 1998).
  - Verb class described by thematic roles (23), selectional restrictions on the arguments, and frames consisting of a syntactic description (55) and semantic predicates (94)
  - 274 Verb Classes and 5257 verb senses
Previous Work (5/5)

Outline

- Motivation
- Previous work
- Language Challenges
- Bengali Subcategorization Frames Acquisition
- Classification of Bengali Verbs
- Evaluation
- Future Task
Language Challenges

- **Properties of Bengali** -
  - Free phrase order language
    
    ```
    Ami kal tomar sathe dekha korbo
    Ami tomar sathe kal dekha korbo
    Ami kal dekha korbo tomar sathe
    Kal ami tomar sathe dekha korbo
    Kal ami dekha korbo tomar sathe
    ```
  
  - Rich morphology (Tense, Aspect and Person for verb)
  
  - Compound verbs ("dekha" (see), "kara" (do), "dekha kara" (meet))
  
  - Difficult to differentiate Arguments from Adjuncts

  ```
  [Mohit] [sakalbela] [kath diye] [bari] [toiri korchilo]
  [Mohit] [was preparing][house] [out of wood ][in the morning]
  ```
Outline

- Motivation
- Previous work
- Language Challenges
- Bengali Verb Subcategorization Frames Acquisition
- Classification of Bengali Verbs
- Evaluation
- Future Task
Bengali Verb Subcategorization Frames Acquisition

- Hypothesis
- Corpus Preparation
- Attempts Done
- Target Verb Identification
- English Equivalent Verb Determination
- VerbNet Frames
- Bengali Verb Subcategorization Frames Acquisition
Hypothesis

- Verb Subcategorization frames of equivalent English verbs (in the same sense) for a Bengali verb
  - Initial set of verb subcategorization frames for that Bengali verb
Corpus Preparation

- Bengali News Corpus (Ekbal and Bandyopadhyay, 2008) developed from web archive
  - 14000 sentences

- POS tagged using Maximum Entropy based POS tagger (Ekbal et al., 2008) (accuracy 88.2%)

- Application of rule-based chunker (accuracy 89.4%) to get the chunked output
Attempts Done (1+10 =11 most frequent Bengali verbs) (1/2)

<table>
<thead>
<tr>
<th>Bengali Verbs</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>তৈরি করা (tai ri kara) (make)</td>
<td>23</td>
</tr>
<tr>
<td>ব্যবহার করা (byabahar kara) (use)</td>
<td>18</td>
</tr>
<tr>
<td>বস করা (bas kara) (live)</td>
<td>17</td>
</tr>
<tr>
<td>কাজ করা (kaj kara) (work)</td>
<td>15</td>
</tr>
<tr>
<td>সংগ্রহ করা (sangraha kara) (collect)</td>
<td>13</td>
</tr>
<tr>
<td>বন্ধ করা (bandha kara) (shut)</td>
<td>13</td>
</tr>
<tr>
<td>নিজেকরা করা (chitkara kara) (shout)</td>
<td>3</td>
</tr>
<tr>
<td>ভুল করা (bhui kara) (mistake)</td>
<td>3</td>
</tr>
<tr>
<td>জিজ্ঞাসা করা (jigya sa kara) (ask)</td>
<td>3</td>
</tr>
<tr>
<td>পর্যবেক্ষণ করা (parjabe khan kara) (observe)</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Verbs</th>
<th>Frequency in the corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>দেখা (dekha) (see)</td>
<td>251</td>
</tr>
<tr>
<td>করা (kara) (do)</td>
<td>206</td>
</tr>
<tr>
<td>আসা (asa) (come)</td>
<td>30</td>
</tr>
<tr>
<td>নেওয়া (noea) (take)</td>
<td>28</td>
</tr>
<tr>
<td>গেরা (ghora) (travel)</td>
<td>19</td>
</tr>
</tbody>
</table>
Most frequent verb  “দেখা” (dekha) (see)

Next highest frequent verb  “করা” (kara) (do)
- A special compound verb in Bengali
Target verb Identification

- Stemming (accuracy 97.09%)
- Retrieve pattern \{(VM)\} for simple verbs and \{[XXX] (NN) [kara] (VM)\} for compound verbs
- Verb forms not considered in the current task
  - Passive forms (e.g. “করানো” karano),
    করিয়ে (kariye)
  - Special forms (e.g. ॐকরক করা (jhakjhak kara),
    তকতক করা (taktak kara))
English Equivalent Verb Determination

- Bengali-English bilingual dictionary
- http://home.uchicago.edu/~cbs2/banglainstruction.html
- Synonymous Verb Set (SVS)

< ব্যবহার করা v. to apply, to use; to behave, to treat...>

< ব্যবহার করা v. SVS1; SVS2; SVS3; ...>
VerbNet Frames (1/3)

- The VerbNet files contain the verbs with their possible subcategory frames and membership information in XML file format.
- The XML files of VerbNet have been preprocessed to build up a general list that contains all members (verbs) and their possible subcategorization frames (primary as well as secondary) information.
- This preprocessed list is searched to acquire the subcategorization frames for each member of the SVS of the Bengali verb.
VerbNet Frames (2/3)

- ..... <VNCLASS ID="use-105"...>
  - <MEMBERS>
    - <MEMBER name="use" wn=""/>
    - <MEMBER name="utilize" wn=""/>
    - <MEMBER name="apply" wn=""/>
    - <MEMBER name="employ" wn=""/>
  - </MEMBERS>
  - </FRAME>
- <DESCRIPTION descriptionNumber="8.1" primary="NP-PP" secondary="for-PP" xtag="0.2"/>
- <EXAMPLES>
  - <EXAMPLE>I spent the money for my training.</EXAMPLE>
- </EXAMPLES>
- </FRAME>

February 12, 2010
<table>
<thead>
<tr>
<th>SVS (VerbNet classes)</th>
<th>Primary and Secondary Frames for a SVS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Make (build-26.1-1)</td>
<td>NP-PP, NP, NP-NP, NP-NP-PP, Asset-PP</td>
</tr>
<tr>
<td></td>
<td>Asset-Subject</td>
</tr>
<tr>
<td>Build (build-26.1-1)</td>
<td></td>
</tr>
<tr>
<td>Use (use-105, consume-66, fit-54.3)</td>
<td>NP-ADVP, NP-PP, NP-TO-INF-VC, Basic Transitive, NP-ING-SC, Location Subject Alternation, NP-PP for-PP, Location-PP</td>
</tr>
<tr>
<td>Apply (use-105)</td>
<td></td>
</tr>
<tr>
<td>Behave (masquerade-29.6, 29.6-1)</td>
<td>PP, Basic Transitive as-PP, like-PP, in-PP</td>
</tr>
</tbody>
</table>
In simple sentences the occurrence of the NNPC, NNP, NNC or NN tags preceded by the PRP (Pronoun), NNP, NNC, NN or NNPC tags and followed by the verb gives similar frame syntax for “Basic Transitive” frame of the VerbNet.

(ami)(PRP) (kakatua)(NNP) (dekhi)(VM)
I parrot see
The syntax of “WHAT-S” frame for a Bengali sentence has been acquired by identifying the sentential complement part of the verb (*dekha*).

The target verb followed by a NP chunk that consists of another main verb and WQ tag (question word) helps to identify the “WHAT-S” kind of frames.

\[(ami)(PRP) (dekhlam)(VM) (NP)((tara) (PP) (ki)(WQ) (korche)(VM))\]

I saw they what did
In order to acquire the frame of “NP-ING-OC”, we have created the list of possible Bengali inflections that can appear for the English “-ING” inflection.

These inflections usually occur in sentences made up of compound verbs with conjunctive participle form (-e) and infinitive form (-te).

If the phrase contains any of these inflections followed by the target verb then it gives a similar description of the VerbNet frame “NP-ING-OC”.

(ami)(PRP) (NP) ((tader) (haste)) (dekhechi)

I them laughing have seen
1 Frame: NP-PP

Max made the hand fan and cover Ram

NP(NN) NP(PRP) CCP(PSP) NP (NN CC NN) VGNF (NN VM)

→ From what Max made the hand fan and cover

2. Frame: Sentential (S)

Ram shouted that he will never come back

NP(NN) VGNF (NN VM) NP (DEM PRP CC NN VM NEG)

→ Ram shouted that he will never come back

February 12, 2010
Outline

- Motivation
- Previous work
- Language Challenges
- Bengali Subcategorization Frames Acquisition
- Classification of Bengali Verbs
- Evaluation
- Future Task
Classification of Bengali Verbs

- Key classes for Key verbs
- Synonyms from Bengali Thesaurus
- Sense based Classification of Synonyms
Key classes for Key verbs

- **Task**
  - Key verb classes (*Cbk*) for each of 11 most frequent verbs (Key verbs)
  - Only 8 key verbs considered in the present study
  - Subcategorization Frames Acquisition for each key verb class

- **Assumption**
  - Synonymous verbs of a key verb with same sense are present in a Key verb Class for the key verb
    → these verbs share same subcategorization frames as their Key verb
Classification of Bengali Verbs

- Key classes for Key verbs
- Synonyms from Bengali Thesaurus
- Sense based Classification of Synonyms
- where “(ক)” indicates the component part “করা”(kara) [do]
- Machine readable Bengali Thesaurus is being developed manually from a printed Bengali Thesaurus
- Entries of only eight key verbs present in Thesaurus
- $Cbk \rightarrow$ Bengali Key class for key verb $Ybk$
- $Cbs \rightarrow$ Bengali Synonymous Class
Classification of Bengali Verbs

- Key classes for Key verbs
- Synonyms from Bengali Thesaurus
- Sense based Classification of Synonyms
Sense based Classification of Synonyms (1/6)

- Key verb and its Bengali synonyms are searched in Bengali to English bilingual dictionary to extract their English equivalent synonyms

- $E^C_k \rightarrow$ English equivalent class of key verb $Y^{bk}_k$
  - $E^C_m \rightarrow$ English equivalent class of a synonymous member verb

$$E^C_k = \{KSVS_1; KSVS_2; \ldots; KSVS_q\}$$
$$E^C_m = \{MSVS_1; MSVS_2; \ldots; MSVS_p\}$$
Example

- Sense of English synonyms (apply, use) is different from the sense (behave) for Bengali key verb ব্যাবহার করা (byabahar kara)

\[
Ybk \to ব্যাবহার করা (byabahar kara) (Cbk=2)
\]
\[
Xbm \to গ্রয়োগ করা (prayog kara)
\]
Sense based Classification of Synonyms (3/6)

- Algorithm
  - If there exists an $Xbm$ such that $Xbm$ belongs to $Cbs$
  - for $i = 1$ to $p$, for $j = 1$ to $q$
  - if $(Zsi \cap Zdj) \neq \emptyset$
  - then $Xbm$ is to be assigned in class $Cb_k$

where $Zsi$ belongs to MSVS and $Zdj$ belongs to KSVS
Sense based Classification of Synonyms (4/6)

# Key Verb:
<बांधना करना v. to apply, to use; to behave, to treat (a person), to behave towards;...

# Member Verbs:
<सोधन करना v. to employ; to apply, to use;>
<सजाना करना v. to behave; to deal (with); to act (towards); to practice;>
Sense based Classification of Synonyms (5/6)

# Key Verb:
< काम करा v. to apply; to use; to behave, to treat (a person) to behave towards; ...>

# Member Verbs:
< काम करा v. to apply; to use;>
< आदेश करा v. to behave; to deal (with), to act (towards); to practice;>
The process terminates when no MSVS of a member verb is left unclassified.
Snapshot

<?xml version="1.0" encoding="UTF-8"?>
<VNCLASS ID="খেখা.xml"
<MEMBERS>
<MEMBER name="খেখা">
<MEMBER name="সম্ভ্রনব "
<MEMBER name="পর্বতকরা "
<MEMBER name="নাগা" 
<MEMBER name="ভাকালা "
</MEMBERS>
<FRAMES>
<FRAME name=Basic-Transitive</FRAME>
<SYN>NP-NP-V</SYN>
<Example>আমি কাকাজুড়া দেখি</Example>
<FRAME name=S</FRAME>
<SYN>NP-V-(NP-NP-V) </SYN>
<Example>আমি দেখলাম যে লাম ঐ কাজটি করছে</Example>
</FRAMES>......
Outline

- Motivation
- Previous work
- Language Challenges
- Bengali Subcategorization Frames Acquisition
- Classification of Bengali Verbs
- Evaluation
- Future Task
Evaluation

- The set of acquired subcategorization frames or the frame lexicon can be evaluated against a gold standard corpus obtained either through manual analysis of corpus data, or from subcategorization frame entries in a large dictionary or from the output of the parser made for that language.
Evaluation (1/2)

- As there is no available parser, we consider mostly simple sentences for the current task.
- Handling the phrase level tagging error caused by chunking.
- 120 correctly chunked sentences prepared manually to make gold standard test data.
Evaluation (2/2)

- Recall (r) - The percentage of subcategorization frame types in the gold standard that the system proposes.
- Precision (p) - The percentage of subcategorization frame types that the system proposes are correct according to the gold standard.
- F-Measure - $2 \times p \times r / (p + r)$

<table>
<thead>
<tr>
<th>Measures</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>70.83%</td>
</tr>
<tr>
<td>Precision</td>
<td>74.11%</td>
</tr>
<tr>
<td>F-Measure</td>
<td>72.44%</td>
</tr>
</tbody>
</table>

Evaluation result
**Results (1/4)**

<table>
<thead>
<tr>
<th>Information</th>
<th>Freq.</th>
<th>Number of KSVSs or Verb Classes of the key verbs after first phase of recursive classification task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sentences in the corpus</td>
<td>14000</td>
<td>15</td>
</tr>
<tr>
<td>Number of key verbs considered in the present task</td>
<td>11</td>
<td>Number of KSVSs or Verb Classes of the key verbs at the end the recursive classification task</td>
</tr>
<tr>
<td>Number of key verb entries available in the Bengali synonyms thesaurus to construct main classes</td>
<td>8</td>
<td>22</td>
</tr>
<tr>
<td>Number of member verbs identified from the Bengali synonyms thesaurus entries</td>
<td>41</td>
<td>Number of subcategorization frames acquired from the chunked gold standard 120 sentences</td>
</tr>
<tr>
<td>Number of member verbs appeared in the corpus with frequency &gt;0</td>
<td>18</td>
<td>85</td>
</tr>
<tr>
<td>Number of sentences containing member verbs in the corpus</td>
<td>120</td>
<td>Number of subcategorization frames identified correctly from the acquired 85 sentences</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Number of newly found subcategorization frames only for Bengali</td>
</tr>
<tr>
<td></td>
<td></td>
<td>63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
</tr>
</tbody>
</table>
### Results (2/4)

| Key Verb | Synonym1 | Synonym2 |...
|----------------|----------------|----------------|
| <Key Verb> | <Synonym1 (Freq1)> | <Synonym2 (Freq2)> |...
| <खेला (dekhā) [see]> | <भजन (takano) {2}>, | <प्रात (chaoa) {2}>, | <प्रसार (darshan kara) {9}>, | <दशक (lakṣya kara) {1}>, | <द्वितीय (dristipat kara) {0}> |
| <दौर (tairi kara) [make]> | <दिनी (nirman kara) {0}>, | <प्रशस्त (prostut kara) {11}>, | <शृंखला (sristi kara) {11}>, | <उत्पादन (utpadon kara) {0} [produce] >, | <स्पष्ट (sthapon kara) {6}>, | <राजस्व (gathan kara) {10}> |
| <व्यय कर (babohar kara) [use/behave]> | <प्रयोग (pryog kara) {7} [apply]>, | <आचरण (achorōn kara) {2} [behave]> | <क्रान्त (bas kara) [live]> | <क्रान्त (bas-bas kara) {5}>, | <अतिरिक्त (adhisthan kara) {0}>, | <अवन्त (abasthan kara) {0}> |

---

February 12, 2010
### Results (3/4)

<table>
<thead>
<tr>
<th>Key Verb</th>
<th>Number of Verb Classes/KSVSs for the Key Verb</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>dekha</code></td>
<td>3</td>
</tr>
<tr>
<td><code>tairi kara</code></td>
<td>3</td>
</tr>
<tr>
<td><code>babahar kara</code></td>
<td>2</td>
</tr>
<tr>
<td><code>kaj kara</code></td>
<td>1</td>
</tr>
<tr>
<td><code>sangroho kara</code></td>
<td>2</td>
</tr>
<tr>
<td><code>chitkar kara</code></td>
<td>1</td>
</tr>
<tr>
<td><code>jigyasa kara</code></td>
<td>2</td>
</tr>
</tbody>
</table>

- **`dekha`**: {काजकरण (takano) [1], चाचाकरण (chaota) [1], देख ण करण (davhan kara) [2], देखी ण करण (lakhyana kara) [2]}
- **`tairi kara`**: {चढ़करण (prostut kara) [1], चढ़ी ण करण (sristi kara) [2], चढ़ी ण करण (sthapana kara) [2], पठल करण (gathan kara) [2]}
- **`babahar kara`**: {प्रयोग करण (pryog kara) [2], आकरण करण (achoron kara) [2]}
- **`kaj kara`**: {पहिचान करण (porishrom kara) [1]}
- **`sangroho kara`**: {प्रयोग करण (jigar kara) [1], बढ़ावण करण (adav kara) [2]}
- **`chitkar kara`**: {हाईचाइ करण (haichai kara) [1]}
- **`jigyasa kara`**: {प्रयोग करण (prokno kara) [1], भिजेस करण (jigves kara) [1], भिज्यापाड करण (jigyasa bad kara) [2]}

February 12, 2010
Results (4/4)

<table>
<thead>
<tr>
<th>Bengali &lt;Key Verb Class&gt;</th>
<th>Type of Subcategory Frames</th>
<th>No. of Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Member Verb)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;দেখা (dekha)&gt; [see]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(ভাকালা)</td>
<td>Basic Transitive</td>
<td>(1)</td>
</tr>
<tr>
<td>(দর্শন করা)</td>
<td>Basic Transitive</td>
<td>(2)</td>
</tr>
<tr>
<td>(লয় করা)</td>
<td>S (Sentential Complement)</td>
<td>(1)</td>
</tr>
</tbody>
</table>

| <বাস করা (bas kara)> [live]       |                            |               |
| (কবর্ষাক করা)                    | Basic Transitive ADVP-PRED | (2)           |
|                                  | For-PP*                    | (1)           |
| <কাজ করা (kaj kara)> [work]      |                            |               |
| (বাণিজ্য করা)                    | NP-PP                      | (1)           |
| <ক্রান্ত করা (sangroho kara)> [collect] |                            |               |
| (বাণিজ্য করা)                    | Transitive (Material object)| (1)           |
| (আদায় করা)                      | PP                         | (2)           |
New Findings

- The sense wise classified verb synsets can be used for the verb entries in Bengali WordNet.
- If a synonym has already been attempted in the Classification process as a key verb, the synonym will not be considered again.
- New Frames possible in Bengali.

Example

(Basic Transitive frame for *prostut kara*)

Ram khabar *prostut korche*

*(Ram is preparing the food)*
Outline

- Motivation
- Previous work
- Language Challenges
- Bengali Subcategorization Frames Acquisition
- Classification of Bengali Verbs
- Evaluation
- Future Task
Future Task (1/2)

- Dependency of bilingual dictionary
- More analysis on Primary and/or Secondary frames
- Error analysis to recover from the loss in precision
- Machine learning Approach
- Thematic roles for Bengali
- Semantic Role Labeling
Future Task (2/2)

- Full-fledged parser for Bengali
- Alignment issues in Machine Translation from English to Bengali
- Argument selection for Question-Answering systems
Thank you
Questions