NOUN AND VERB GROUP IDENTIFICATION FOR HINDI

- *Smriti Singh, Om P. Damani, Vaijayanthi M. Sarma*

COLING 2012
DEC 14, 2012

Outline

- Introduction - Word Group Identification
- Need for Word Group Identification
- Major Contribution
  - In-depth structural analysis of Noun and Verb Group constituents
  - Procedure for NG and VG identification
- Implementation of Group Identification in Hindi POS Tagger
- Performance Evaluation

Chunking or Word Grouping

Word Group Identification (Chunking)

- No unique definition (some type of shallow parsing):
  - *Chunk: truncated versions of phrase-structure grammar phrases without arguments or adjuncts* (Grover and Tobin 2006)
  - *Chunking identifies major constituents of a sentence without further identifying a hierarchical structure that connects and arranges the chunks* (Abney 1991)
- Chunk: a Head node and its modifiers:
  - [The tall man] [was sitting] [on his suitcase]
  - [ləmbā ādmī] [apne sandūk pe] [baithā thā]
- Non-recursive: Only one head of a lexical category in a chunk:
  - [Ram's] [son]
  - [raam kā] [betā]
- Chunks do not include complements unlike phrases
  - Phrases: [usne]NP [[khānā]NP khāyā]VP
  - [usne]NG [khānā]NG [khāyā]VG

Motivation for Word Group Analysis

- Resolve PoS ambiguities
- Help in next level of parsing

Word Group Identification Process

- Structural Analysis
- Morphotactical information
- Part-of-Speech (POS) Information
  - Interplay between POS Tagging and Group Identification

Motivation for Noun Group Identification in Hindi

1. To deal with the ambiguities between
   - Demonstrative and Personal pronoun
   - Adjective and noun
   - Ordinal and noun
   - Noun and verb

Motivation for Noun Group Identification

- Demonstrative and Pronoun ambiguity

  us ***kāl-e*** *ghoɾ-e* *ko* *rok-o*

  that-obl *black-obl* *horse-obl* *ACC* *stop-imp*

  'stop that **black** horse'

  Desired tagged output for the five words:

  Demonstrative Adjective Noun Postposition Main Verb

  Tagger's incorrect output:

  Pronoun Adjective Noun Postposition Main Verb

Motivation for Noun Group Identification

- Adjective and Noun ambiguity

As ADJ: *əcch-e* ***kām*** *kā nətijā əcchā nikəl-t-ā* *hai*

good-obl ***deed*** *of result good turn-hab,masc,sg be-pres*

'Do good have good'

As NOUN: *əcch-e* *kā nətijā əcchā nikəl-t-ā* *hai*

good-obl of result good turn-hab,masc,sg be-pres

'Do good have good'

Motivation for Noun Group Identification

- Ordinal and Noun ambiguity

As ORD: *dūsr-e* ***lə[k-e*** *ne* *kəh-ā*

second-obl ***boy****-obl ERG say-perf*

‘the second boy said’

As NOUN: *dūsr-e* *ne* *kəh-ā*

second *ERG* *say-perf*

‘the second said’

Motivation for Noun Group Identification

- Verb and Noun ambiguity

As Noun: *tair-**n-e** ke bəhut lābh haĩ*

swim-Inf-obl Poss many benefits be-pres,pl

‘Swimming has many benefits’

As Verb: *tair-**nā** bəhut lābhkārī hai*

swim-Inf very beneficial be-pres

‘Swimming is very beneficial’

Categorization of NG Constituents

NGs are formed around a noun/pronoun that acts as a nucleus in the group preceded by many pre-nominal categories

NG = (Set 1)* (Set 2)* (Set 3)* Set 4 (Set 5) (Particle)

Some examples of Hindi NGs:

- Dem Pron+N, e.g., *vo mez (that table)*
- Poss pronoun+N, e.g., *merā kəmrā (my room)*
- Adj+N, e.g., *sundər ləɽkī (beautiful girl)*
- Dem pron+Adj+N, e.g., *vo sundər ləɽkī (that beautiful girl)*
- Card+N, e.g., *cār ghoɽe (four horses)*
- Ord+N, e.g., *dūsrā ləɽkā (second boy)*

Categorization of NG Constituents

Set 1 includes:

- Demonstrative Pronoun
- Possessive Pronoun

Ordering:

((Demonstrative) (Possessive)) OR ((Possessive) (Demonstrative))

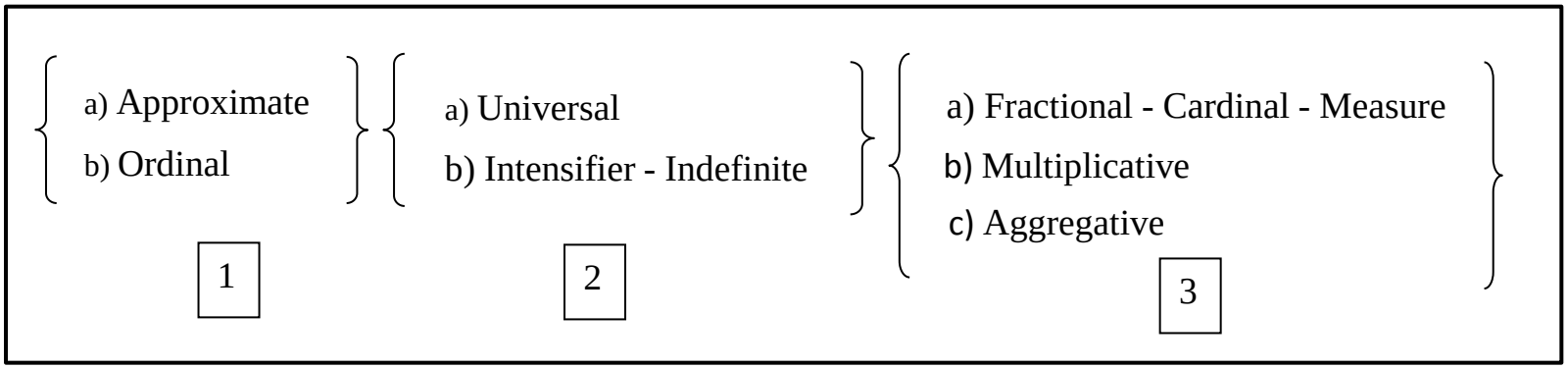
- The ordering suggests that any of the following outputs are valid:
- Both are optional – (*vo tumhārī) mīthī bātẽ*
- Both may appear together - *vo tumhārī mīthī bātẽ* or *tumhārī vo mīthī bātẽ*
- One may appear without the other - *vo mīthī bātẽ* or *tumhārī mīthī bātẽ*

NG = (Set 1)* (Set 2)* (Set 3)* Set 4 (Set 5) (Particle)

Categorization of NG Constituents

Set 2 includes: Intensifiers, Numerals (approximate, fractional, ….)

Ordering among the constituents:



bəhut kəm log 'very few people' / dugunā ləmbā rāstā 'double long distance'
kuch zyādā log 'few more people' / ləgbhəg prətyek vyəkti 'almost every person'

NG = (Set 1)* (Set 2)* (Set 3)* Set 4 (Set 5) (Particle)

Dem/Poss

[un səbhī]

Categorization of NG Constituents

Set 3 includes:

Adjectives

Ordering:

- ((Verbal Adjective) (Adjective))

The order generally followed by different kinds of adjectives is quality-size-age-shape-color-origin material, for example:

ləmbī kālī reshmī bənārəsī sāṛī (long black silk banarasi saree)

nəyā khushhāl bhārtiyə səmudāyə (new happy Indian community)

NG = (Set 1)* (Set 2)* (Set 3)* Set 4 (Set 5) (Particle)

Dem/Poss Int/Num

[un səbhī yuvā sərkārī]

Categorization of NG Constituents

Set 4: Heads

- Right-most elements of the group (exceptions- postpositions and particles)
  - Noun
  - Proper Noun
  - Gerund
  - Pronouns (except demonstrative and possessive pronoun)

NG = (Set 1)* (Set 2)* (Set 3)* Set 4 (Set 5) (Particle)

Dem/Poss Int/Num Adj

[un səbhī yuvā sərkārī kərəmchārīyo]

Categorization of NG Constituents

Set 5: Postpositions

- Primary (ne, ko, ke,)
- Compound (ke bāḍ’, ‘ke sāṭʰ)

NG = (Set 1)* (Set 2)* (Set 3)* Set 4 (Set 5) (Particle)

Dem/Poss Int/Num Adj

[un səbhī yuvā sərkārī kərəmchārīyo ko]

Categorization of NG Constituents

- **Particles or discourse markers** may appear at many places

  hī (only), bhī (also/too), to (at least), tək (even), bhər (all)

  - *ek hī kitāb lānā*
    one only book get
    'Get only one book'
  - *ek kitāb hī lānā*
    one book only get
    'Only get a book'
  - *ek kitāb bhī lānā*
    one book also get
    'Also get a book'

NG = (Set 1)* (Set 2)* (Set 3)* Set 4 (Set 5) (*Particle*)

Dem/Poss Int/Num Adj Head Postp

[un səbhī yuvā sərkārī kərəmchārīyo ko bhī]

Ordering of NG Constituents

NG = (Set 1)* (Set 2)* (Set 3)* Set 4 (*Set 5*) (Particle)

[un səbhī yuvā sərkārī kərəmchārīyo ko bhī] chhuttī pər haĩ
Those your all young government employees too leave on be-pres,pl
'All those government employees of yours too are on leave'

Computational Rules for NG Identification

1. For all tokens, processing goes from right to left

1a. Look for a Set 5 or a Set 4 element to start an NG

1b. If Set 5 member, i.e., a postposition is found

1b (i) Oblique NG has started

1c. If Set 4 element is found

1c (i) Direct NG has started

1d. If a Demonstrative pronoun is found

1d (i) Consider it as a Pronoun (head)

NG = (Set 1)* (Set 2)* (Set 3)* Set 4 (Set 5) (*Particle*)

Dem/Poss Int/Num Adj Head Postp

[un səbhī yuvā sərkārī kərəmchārīyo ko bhī]

Computational Rules to Identify an NG

2. *If oblique NG has just started with a Set 5 element, i.e., with a postposition*

2a. Look for a Set 4 element

2b. If Set 4 element is not found; find the list of possible POS tags for the current word

2c. If a POS Tag appears in the possible POS Tags' list and also in Set 4

2c (i) Assign the tag which is common to both.

2d. If there is no common element in the list and Set 4s

2d (i) Assign the tag other than PP to the next word using the list of possible tags for it.

NG = (Set 1)* (Set 2)* (Set 3)* Set 4 (Set 5) (*Particle*)

Dem/Poss Int/Num Adj Head Postp

[un səbhī yuvā sərkārī kərəmchārīyo ko bhī]

Computational Rules to Identify an NG

- *If any NG has started*

3a. Look for a Set 3 and/or Set 2 and/or Set 1 element

3b. If Set 3, 2 and 1 elements are found

3b (i) The NG includes the current word

3c. If set 3, 2 and/or 1 elements are not found

3c (i) The NG has already ended with the previous word

4. If any NG is completely identified

4a. Apply rules to check the agreement between modifiers/qualifiers and their head and do corrections if necessary

5. Start looking for the next NG

NG = (Set 1)* (Set 2)* (Set 3)* Set 4 (Set 5) (*Particle*)

Dem/Poss Int/Num Adj Head Postp

[un səbhī yuvā sərkārī kərəmchārīyo ko bhī]

Rule Application in NGI

Step by step application of rules from right to left:

a) *vo kāl-e ghod-e ko rok rəh-ā hai*
he black-obl horse-obl ACC stop prog-masc,sg be-pres
'He is stopping the black horse'

- Start scanning the sentence from right to left
- Found *'ko'* (assume Oblique NG has started) – SET 5 element
- Found a Noun *ghod-e* which is in oblique form – SET 4 element
- *kāl-e* found as a qualifier and is also case-marked – SET 3 element
- Include *kāl-e* in the NG as case, gender and number features match
- Found *vo* that can be a pronoun or demonstrative – SET 1 element
  - Reject 'demonstrative' as *vo* does not agree with the head noun for oblique case
  - Reject DEM and Tag *vo* as PRON

Rule Application in NGI

Step by step application of rules from right to left:

b) *vo kālā ghodā so rəh-ā hai*
that black horse sleep prog-masc,sg be-pres
'That black horse is sleeping'

- Start scanning the sentence from right to left
- Found a Noun *ghodā* which is in direct case - SET 4 element
- *kālā* found as a qualifier and is not case-marked – SET 3 element
- Include *kālā* in the NG as case, gender and number features match
- Found *vo* that can be a pronoun or demonstrative – SET 1 element
  - Reject 'pronoun' as *vo* as there cannot be two heads in an NG
  - Accept DEM and Tag *vo* as DEM

Motivation for Verb Group Identification

To solve the ambiguities between:

- Main Verb and Verb Auxiliary

  [rəh rəh-ā hai]

  live prog-masc,sg be-pres

  'is living'

- *'rəh'* is ambiguous as it may be Main Verb or Auxiliary Verb
- Auxiliaries appearing after *rəh* may help resolve the ambiguity

Motivation for Verb Group Identification

To solve the ambiguities between:

- Main Verb and Noun

[kər cuk-ā thā]

do comp-masc,sg be-past

'had done'

System may use the information that *cuk* as an auxiliary followed by a tense auxiliary requires a main verb to precede it. This information rules out the Noun tag and leaves Main Verb as the correct tag.

Motivation for Verb Group Identification

To solve the ambiguities between:

- Main Verb and Noun

kər [cuk-ā de-g-ā]

tax pay-masc,sg give-fut-masc,sg

'will pay the tax'

The system may consider *kər* to be a part of the VG and will output the VG as *kər cukā degā*. Thorough analysis and strict morphotactical rules help choose the correct option in such constructions. A constraint that says that the completive aspectual auxiliary *cuk* cannot be followed by the modal auxiliary *de* needs to be applied in order to resolve the ambiguity.

Motivation for Verb Group Identification

- Suffixes may be ambiguous
  - Conditional mood and habitual aspect marker

*bādəl roz [ā-**t-e** the]*

Clouds everyday come-hab be-past
'Clouds would come everyday'

*əgər bādəl roz [ā-**t-e**]*

if clouds everyday come-cond

'if clouds came everyday'

Motivation for Verb Group Identification

- Feature Agreement is needed to resolve ambiguities

  vo merā bhāī thā

  he my brother be-past

  'he was my brother'

  **'bhāī thā'*

  like-past-fem be-past-masc

  was liked'

- *bhāī* is ambiguous for the tags Verb and Noun
- As a Verb, gender of *bhāī (fem)* and of the tense auxiliary '*thā' (masc)* mismatch
- Verb tag is rejected and Noun is chosen

Constituents of a Hindi VG and their Order

Start Marker	Intermediate Markers		Must End Markers
	Possible End Markers	Must-Continue Markers	
Main Verb (Root)	Necessity Perfective-gen-num Subjunctive-per-num	Ability/Probability, Obligation/Permission Habitual/Progressive Perfective Passive Infinitive	Present Tense Past Tense Future+gen-num Imperative Conditional-gen-num

- Basic Order:

  Verb Root−Infinitive/Passive−Modal Auxiliary−Aspect−Tense−Mood

- A VG is identified by scanning the sentence from left to right using the expression:

  Start Marker (Intermediate marker)* Must-end marker

Constituents of a Hindi Verb Group

1) Start Markers: Main Verb

2) Intermediate Markers:

a. Possible end markers:

- Modal Auxiliary: *(cāhie)* 'should'
- Aspect: - *(-yā),* - *(-ā),* - *(-ā),* - *(-ī),* - *(-e),* - *(-e),* - *(-ī),* - *(-ĩ),* - *(-ĩ)*
- Subjunctive: - *(-ũ),* - *(-ũ),* - *(-e),* - *(-e),* - *(-ŋ),* - *(-ẽ),* - *(-ẽ),* - *(-o)* ,- *(-o)*

b. Must Continue Markers:

- Aspect: Habitual - *(-t)*, Progressive *(rəh)*, Completive *(cuk)*
- Modal Auxiliaries: Ability/probability: *(sək), a*bility: *(pā), o*bligation: *(pəɽ)*, *p*ermission: (*de)*
- Passive: Perfective marker followed by the passive marker *jā,* e.g., / /

3) Must-end Markers

- Future with gender-number: - *(-gā),* - *(-gī),* - *(-ge)*
- Imperative mood: *null,* - *(-o)* ,- *(-o), (-ie), (-ie), (-jie),* – *(-nā)*
- Tense Auxiliary: Present: *(hai), (haĩ),* Past: *(thā), (the), (thī), (thĩ)*
- Conditional Mood marker - - *(-t-)*

Procedure for VG Identification

Hindi VGs are identified by scanning the sentence from left to right using the expression:

Start Marker (Intermediate marker)* Must-end marker

- Start-marker and must-end markers are obligatory
- Intermediate markers are optional and may recurse (marked as *)

Performance Evaluation with a CRF based Hindi POS Tagger

Experiment	Average Accuracy of 4 folds
Only CRF	95.18%
CRF + NGI after	95.67%
CRF + VGI after	95.73%
CRF + NGI after + VGI before	95.87%
CRF + NGI after + VGI after	95.26%

- Both NGI and VGI help improve accuracy
- Best performance obtained with VGI applied before CRF and NGI after CRF
- Error reduction of major POS categories is 15% (from 4.72% to 4.1%)
- Last 5% errors remain due to
  - Corpus inaccuracies
  - Annotators disagreement
  - Long-distance dependencies
  - Non-handling of Compounds

Standing Challenges/Problems

Verb-Noun Ambiguity

- *mætʃ 48-48 ovərõ kā kər diyā gəyā hai*

  match 48-48 overs of do has been be-pres

  'Match has been made of 48-48 overs'

  (*'diyā gəyā hai*' identified as VG and '*kər*' is marked as a verb (do))

- *mætʃ kā kər diyā gəyā hai*

  match of tax give-past has been

  'Tax has been given/paid for the match'

  (*'diyā gəyā hai*' identified as VG and '*kər*' is marked as a noun (tax))

('*kər*' appears in the same context in the two sentences; difficult to disambiguate without sentence level analysis/subject-object information)

Standing Challenges/Problems

- Proper Name ambiguity with other POS categories

  tīm ne spænish līg lā līg kā khitāb jītā
  Team-ERG Spanish League Lā Liga of prize win-past
  'The team won the Spanish League La Liga title'

  (*'La Liga'* is a proper name but morphological analysis (*lā*) calls it a verb as it is valid verb root. In absence of a sophisticated Proper Noun identification system, Tagger chooses Verb as an appropriate tag)

Problems

- The System does not handle cases of scrambling
  - *tum kya dekh rahe ho?* – ‘dekh rahe ho’ as VG
  - *tum dekh kya rahe ho?* – ‘rahe ho’ as VG

  (rules to handle scrambling are still not in place)

- May lead to faulty grouping in some cases
  - *un-kī yojnāẽ shāntipūrnə uddeshy-õ ke [**liye** haĩ]*
    Their plan peaceful aims-obl of be-pres
    ‘Their plan is for peaceful aims’

  - *un-kī yojnā shāntipūrnə uddeshy-õ ke **liye** [hai]*
    Their plan peaceful aims-obl for be-pres

  (‘*liye*’ preceded by ‘*ke*’ appears more as Verb rather than Postposition in most of the sentences in the learning data)

Future Directions

- Incorporate Proper Noun Identification Rules
- Incorporate Compounds and Conjuncts Identification rules
- Handle cases of scrambling
- Add more learning data to avoid sparsity and reduce ambiguity
- Play more with VGI and NGI's position in the system to get the best performance

Thank You!