POS Tagger for Hindi

Jayesh Suryavanshi

Shashwat Babhulgaonkar

Vivek Sanap

Under the guidance of:

Prof. Smita T. Patil

Outline

- Motivation
- Introduction
- Hindi POS Tagger
- Challenges
- Stages

Motivation

- Part-of-Speech (POS) tagger is the basic building block for various NLP tools
- Wide applications
  - Information Retrieval, Machine Translation, Word Sense Disambiguation, Question Answering System etc.
- Efficient POS tagger has not been reported for Hindi

Introduction

- POS tagging is the process of identifying lexical category of a word on the basis of its context in the sentence

  Input : राम खेल रहा है .

  Output : राम_[PPN] खेल_[VM_MSX_PrDX] रहा_[VAUX]
  है_[VAUX] ._[.]

  (PPN: Proper noun, VM_MSX_PrDX: Verb main (male, singular, present, durative), VAUX: Verb auxiliary)

- Classification
  - Rule- based, Stochastic and Hybrid
  - Supervised and Unsupervised

Hindi POS Tagger

- Rule-based tagger
  - Hindi morphologically rich
    - Morphological analysis helps in
      - Determining the category
      - Determining the feature value
- Uses manually formulated rules at various stages
  - Non-availability of tagged corpora

Challenges: Hindi POS Tagging

- Morphological Analysis
  - Determining category and values of feature (gender, number, person, etc.) from morphemes present in word
- Resolving ambiguities
  - Multiple suffix: “खेलता” -> “ता” or “ ा”
  - Multiple category: “चमकता” -> verb or adjective
  - Multiple feature values: “लड़के” -> singular oblique or plural direct
- Handling unknown words
  - Foreign word (गुडबाय), Proper noun (सलमान), Spelling mistake, etc.

Resources and Stages

RESOURCES

- Lexicon
- Suffix-replacement rules
- Unique suffix list
- Derivational morphology rules
- Suffix analysis
- Stem analysis
- Morpheme flag map
- Multi-category disambiguation rules
- Verb-group analysis rules
- Multi-analysis disambiguation rules

STAGES IN TAGGING

- Tokenisation
- Stemming
- Morpheme analysis and flagging
- Multi-category disambiguation
- Verb-group identification
- Phrase level analysis
- Tag generation

Cleaning and Tokenisation

- Separating special characters attached to words
  - Input: “मैं घर जा रहा हूँ”
  - After cleaning: “ मैं घर जा रहा हूँ ”
- Sentencification: Identifying sentences
  - Input: राम अच्छा लडका है। वह सबका आदर करता हैं।
  - After cleaning:
    - Sentence 1: राम अच्छा लडका है।
    - Sentence 2: वह सबका आदर करता हैं।
- Tokenisation: Breaking into units processed by the system
  - Input: “मैं घर जा रहा हूँ”।
  - Tokens: “, मैं, घर, जा, रहा, हूँ , ”, ।

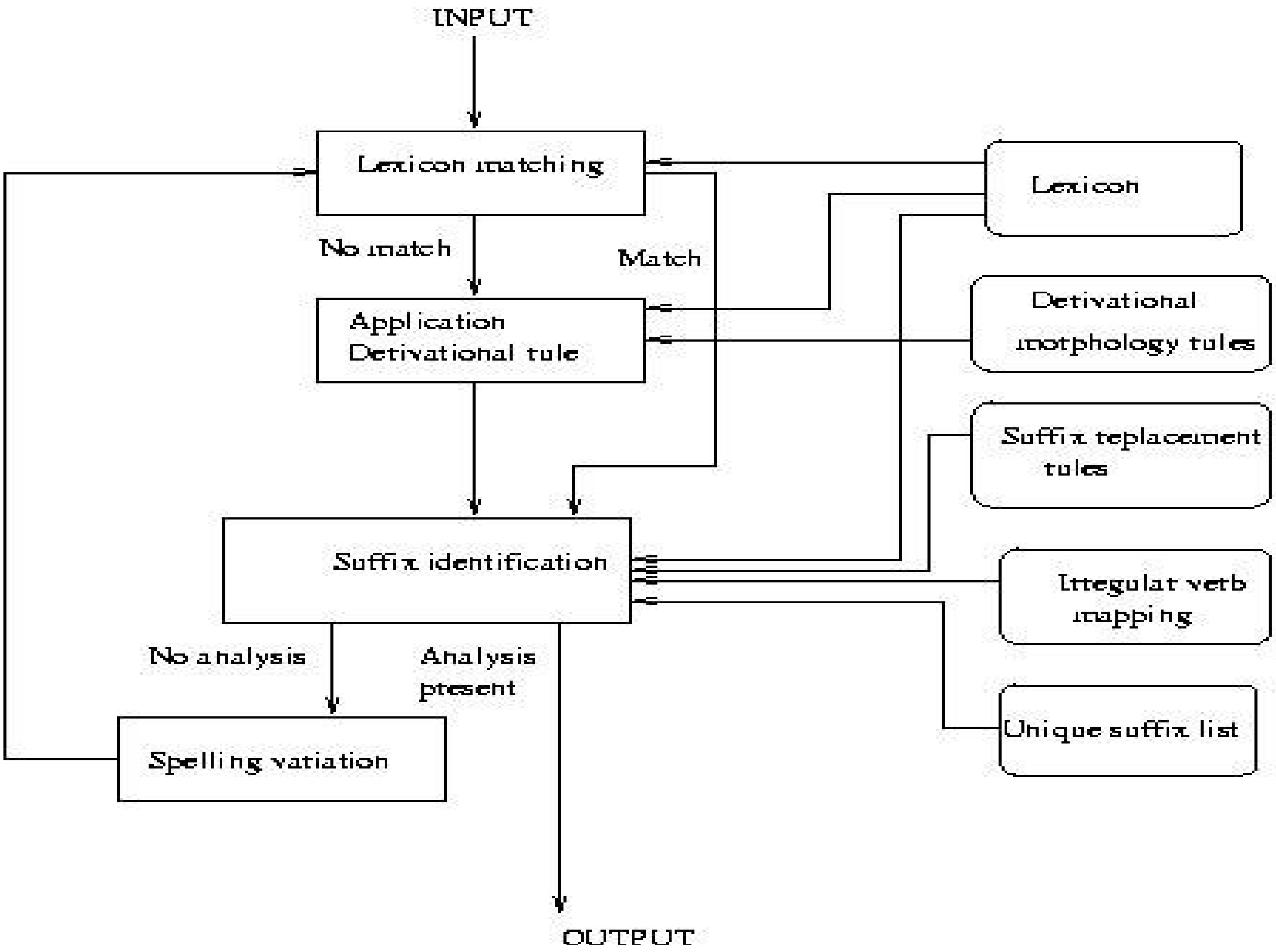
Morphological Analyser

- Identifies and analyses the structural component of the word
- Involves two stages
    - Suffix and category identification by stemmer
    - Analysis by morpheme analyser
- Applications: WorldNet API's, aAQUA search engine

Stemmer

- Provides
  - Stem, suffix and grammatical category
    - Input word: लड़कों
    - Output
      - Stem : लड़का
      - Suffix : ों
      - Grammatical Category : Noun
- Performs initial tagging
  - Output all possible categories for input word
    - Input word: चमकता
    - Output categories: Verb, Adjective
- Heuristics for handling unknown applied at this level

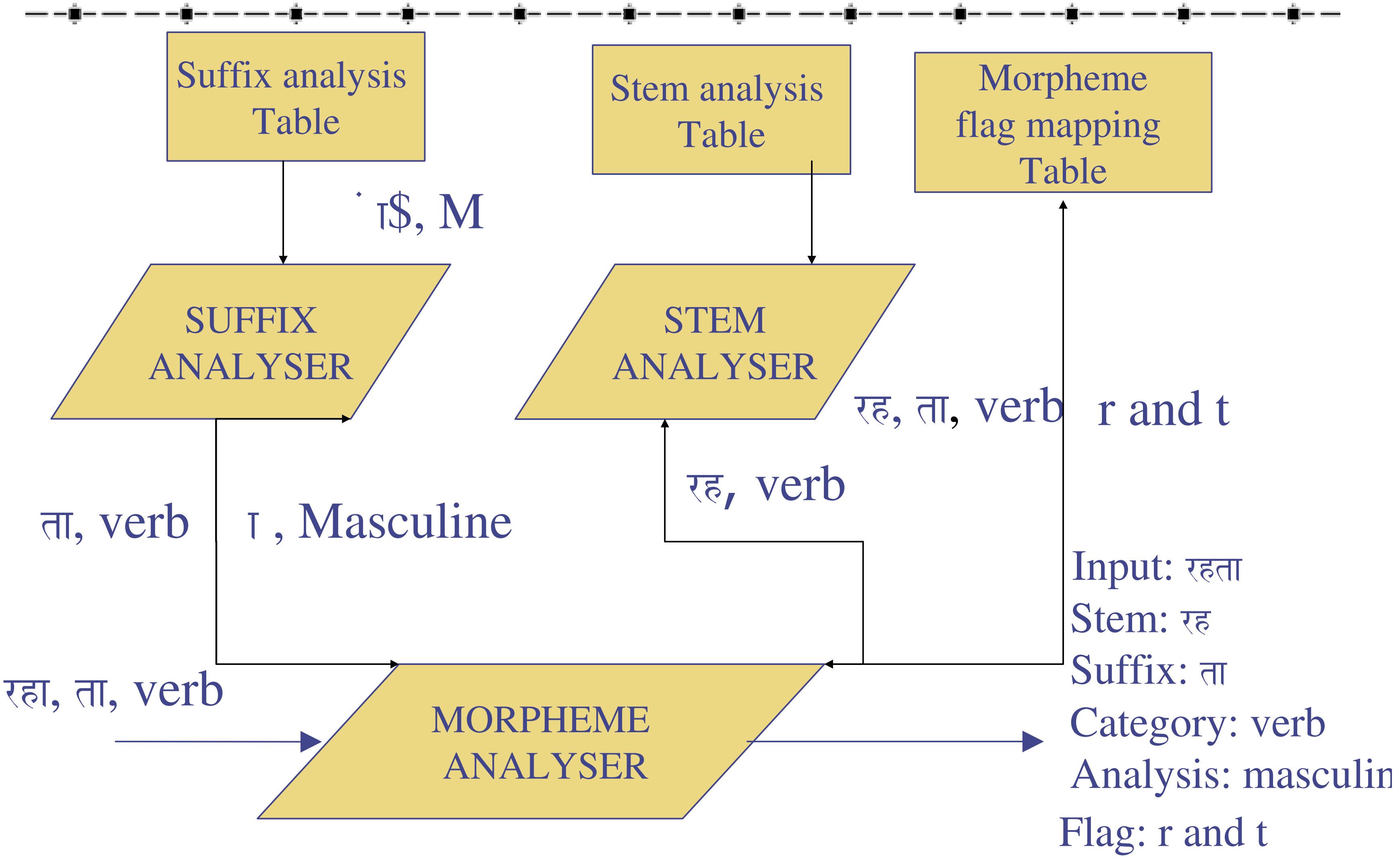
Stemmer Block Diagram



Morpheme Analyser

- Provides grammatical information for the word from the constituent morphemes
  - Verb: gender, number, person, tense, aspect and mood
  - Noun: number, case
  - Pronoun: number, person
- Involves Stem analysis and Suffix analysis
- Flags the presence of morpheme in suffix and stem
  - Used for phrase level analysis of verb

Morpheme Analyser Block Diagram



Suffix analysis Table

Stem analysis Table

Morpheme flag mapping Table

ा$, M

SUFFIX ANALYSER

STEM ANALYSER

रह, ता, verb

r and t

रह, verb

ता, verb

ा , Masculine

रहा, ता, verb

MORPHEME ANALYSER

Input: रहता
Stem: रह
Suffix: ता
Category: verb
Analysis: masculin
Flag: r and t

Morpheme flagging

- Flag the presence of morpheme
  - Used for verb-group analysis
  - Uses morpheme-flag map table
  - Example
    - Input word: “रहता”
    - Flags present: r for ‘रह’ and t for ‘त’

Multiple Category Disambiguator

- A word can occur in multiple categories
  - “खेल” can be verb and noun
- Results show 25% (approx) of the words get multiple categories
- Manually formulated disambiguation rules are used
- At present system is using 32 rules
- 30% of ambiguous words gets disambiguated using these rules

Multiple Category Disambiguation Rules

- Rule format
  - **PRESENTCAT <pcat> CONTEXT-INFORMATION <ntag> THEN <ctag>**
  - CONTEXT-INFORMATION can be like
    - NEXTTAG – next word’s tag
    - PREVIOUSTAG – previous word’s tag

- Rule: PRESENTCAT adverb,adjective NEXTCAT verb THEN adverb
  - Before applying rule: “दोस्ती_[N_S_X] को _[CM] लगातार _[ADJ ADV] बढ़ाना_[VM_MXX_NXX] है _[VAUX]।”
  - After applying rule: “दोस्ती_[N_S_X] को _[CM] लगातार _[ADV] बढ़ाना_[VM_MXX_NXX] है _[VAUX]।”

Verb-Group Identification

- Verb-group comprises finite main-verb and its auxiliaries
  - Example:
    - Input sentence: “राम खेलता रहता है।”
- Useful for
  - Main verb identification, “खेलता रहता है।” , “घर में रहता है।”
  - Aspect & Mood information
- Identification needs determining category
  - Mark the beginning of verb- group, e.g. verb
  - Mark the end of verb- group, e.g. copular verb
  - Come between in verb- group, e.g. neg, particle

Phrase Level Analysis

- Uses context information of word for analysis
- Task performed at this level:
  - Verb group analysis
    - Identifying aspect and mood information
  - Multiple analysis disambiguation

Verb-group Analysis

- Use rules and morpheme flag information for analysis
  - Verb PRESENTFLAG <pflag>CONTEXT-INFORMATION <nflag>THEN <ana>
  - CONTEXT-INFORMATION can be like
    - NEXTFLAG (Flag of word next to main-verb)
    - NEXTFLAG2 (Flag of word 2 positions ahead of main-verb)
- Example,
  - Input sentence: “राम खेलता रहता है। ”
  - Verb-group: “खेलता रहता है”
  - Flags: “खेलता” - t, “रहता” – rt, “है” – null
  - Rule applied: verb PRESENTFLAG t NEXTFLAG rt THEN A:H
  - Analysis: Aspect Habitual (H) in verb group

Multiple Analysis Disambiguation

- Multiple analyses of morpheme in suffix is possible, e.g “ े” of “ लड़के ”
  - “ े” in “ लड़के खेल रहे है” provides plural direct information
  - “ े ” in “ लड़के ने अच्छा खेला ” provides singular oblique information
- Disambiguation with the help of rules
  - noun NEXTCAT cm THEN N:S,C:O
    - Means If the noun has multiple feature value and the category of next word is case-marker (cm) then the correct analysis of noun is singular number and oblique case

Tag Generation

- Category and feature value information is presented in the form of tag
- Properties of Tagset
  - Broad coverage: Tags for major categories
  - Readability: Fixed tag format for categories with feature values,
    - Verb -> VM_GNP_TAM
    - Noun -> N_N_C
- At present tags for 17 categories excluding categories with feature values
- Number of tags including categories with feature values expected to be greater then 500