Center For Indian Language Technology

Indian Institute of Technology, Bombay

Summer Internship Presentation
(May-July) 2014

Submitted By :
Harshita Jhavar
Final Year CSE Undergraduate
MANIT Bhopal

Under The Guidance Of :
Dr. Pushpak Bhattacharyya
Dr. Manish Srivastava

Presentation Outline

- Research Problem Statement
- Motivation
- Approach
- Results
- References
- Future Scope
- Internship Work Timeline

Problem Statement-

“IdentifyingThe Gender Category Of Nouns (Common/Proper) In Hindi Corpus”

Eg. - राम, जो मिठाई बेचता है, आज घर चला गया |

M F M

MOTIVATION:

>Gender of proper and common nouns cannot be identified unless they are specifically assigned a gender in the predefined lexicon list

>It's tedious to make a list of all the existing proper nouns and manually assign their gender category to them

>Failure of precise morphology analysis of nouns leads to overall decrease in efficiency of the morphology analyzer techniques and loss of information related to attributes of existing nouns

Can Gender Category be recognised at Morphology / POS Tagging Level ?

Ending Sound /
Suffix (If Any)(If Root Form Can Be Identified)

>>Dependent on the root word list
And suffix replacement rules

>>Ending Sound (Not Always Right)
अ - पुस्तक, छत(f) , घर, हाथ (m)
ई - लड़की (f) , माली (f)

>>Not All nouns have suffixes

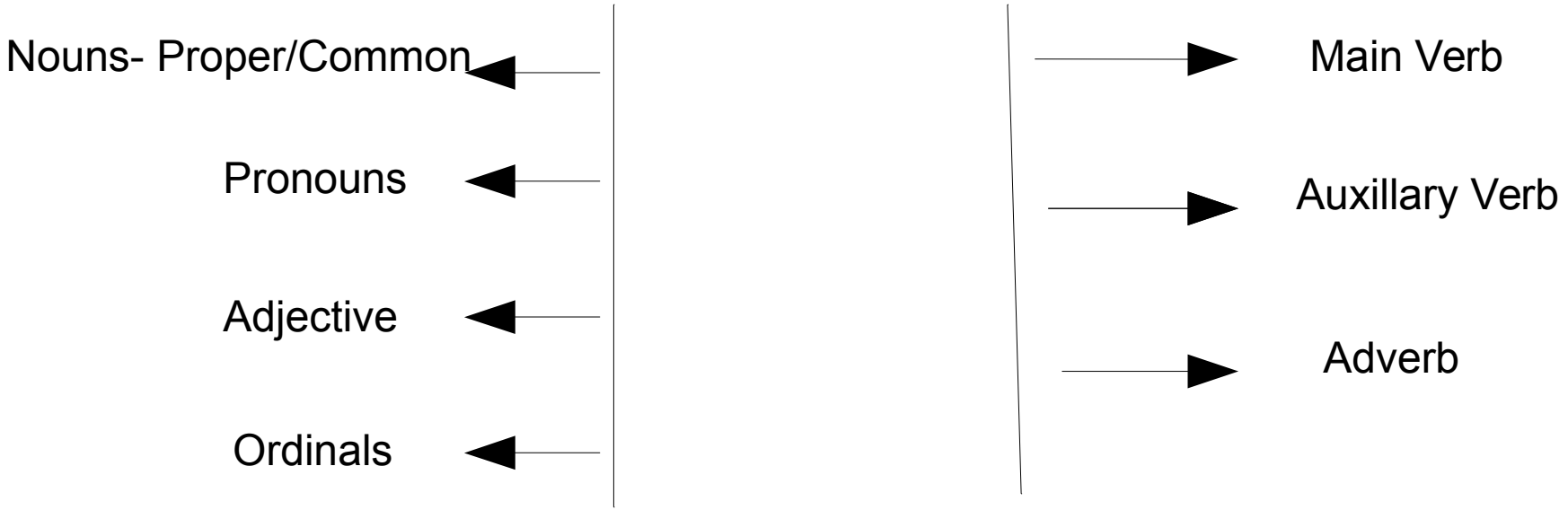
Number
(For Common Nouns)

>> Word Form Along With Plurals.
लड़का - लड़कें
क्रीड़ा - क्रीडायें
But ऋतु, ओं - feminine
शत्रु, ओं - masculine
>> Also plural to singular is possible but not vice versa

Gender of Proper Nouns / Common Nouns cannot be identified at these levels with notable efficiency.

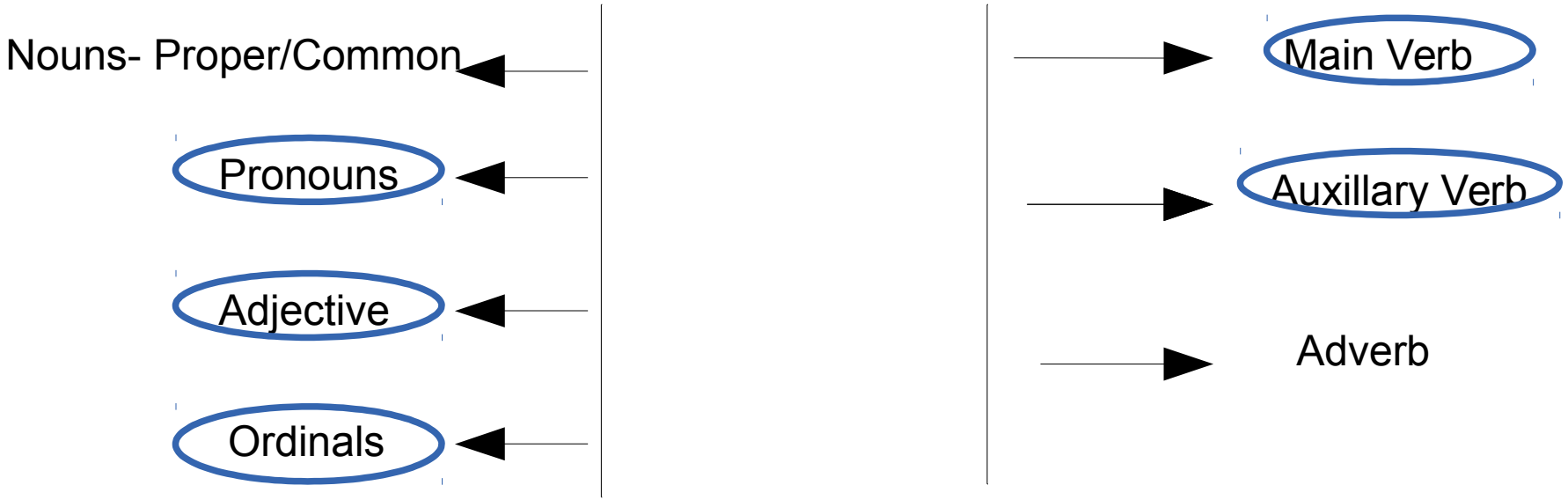
Taking The Approach To Next Level Of Chunking And Parsing ----

Sentence = Noun Phrase + Verb Phrase



Taking The Approach To Next Level Of Chunking And Parsing ----

Sentence = Noun Phrase + Verb Phrase



--- Categories which change according to gender

Machine was Trained For Gender Identification Of **Nouns** via Stanford Classifier 3.4 using The Following Features:-

1.useNGrams=true
1.usePrefixSuffixNGrams=true
1.maxNGramLeng=10
1.minNGramLeng=1
1.binnedLengths=10,20,30
2.useString=true
3.usePrefixSuffixNGrams=true
4.useString=true
5.realValued=true
6.useString=true
7.useString=true
8.useNGrams=true
8.usePrefixSuffixNGrams=true
8.maxNGramLeng=6
8.minNGramLeng=1
9.useString=true

m	उत्पादन	n	NP	head	3	_	उत्पादन	0_में	sg
_	में	psp	NP	child	1	_	में	_	_
_	आई	v	VGNF	head	4	f	आ	या	sg
m	गिरावट	n	NP2	head	17	_	गिरावट	0_का_वजह_से	sg
_	की	psp	NP2	child	4	f	का	_	_
f	वजह	n	NP2	child	4	_	वजह	_	_
_	से	psp	NP2	child	4	_	से	_	_
m	गेहूं	n	NP3	head	10	_	गेहूँ	0_का	sg
_	की	psp	NP3	child	8	f	का	_	sg
f	कीमत	n	NP4	head	17	_	कीमत	0_में	sg
_	में	psp	NP4	child	10	_	में	_	_
m	उछाल	n	NP5	head	14	_	उछाल	0_का	sg
_	की	psp	NP5	child	12	f	का	_	pl
f	आशंकाओं	n	NP6	head	17	_	आशंका	0_को	pl
_	को	psp	NP6	child	14	_	को	_	_
_	खारिज	adj	JJP	head	17	any	खारिज	_	any
_	करते	v	VGNF2	head	21	m	कर	ता_हो+या	sg
_	हुए	v	VGNF2	child	17	m	हो	या	sg
f	सरकार	n	NP7	head	21	_	सरकार	0_ने	sg
_	ने	psp	NP7	child	19	_	ने	_	_
_	कहा	v	VGF	head	0	m	कह	या_है	sg
_	है	v	VGF	child	21	any	है	है	sg
_	कि	avy	CCP	head	21	_	कि	_	_
_	वर्तमान	adj	NP8	child	25	any	वर्तमान	_	any
f	जरूरतों	n	NP8	head	28	_	जरूरत	0_को	pl
_	को	psp	NP8	child	25	_	को	_	_
_	पूरा	adj	JJP2	head	28	m	पूरा	_	sg
_	करने	v	VGNN	head	37	any	कर	ना_के_लिए	any
_	के	psp	VGNN	child	28	_	के	_	_
_	लिए	psp	VGNN	child	28	_	लिए	_	_
m	देश	n	NP9	head	37	_	देश	0_में	sg
_	में	psp	NP9	child	31	_	में	_	_
m	खाद्यान्नों	n	NP10	head	36	_	खाद्यान्न	0_का	pl
_	का	psp	NP10	child	33	m	का	_	sg

SAMPLE TRAINING DATA

RESULTS OBTAINED

Corpus Identity	No. Of Lines		Accuracy For Nouns
	Training Set	Testing Set	
Hindi Tree Bank – LTRC, IIIT Hyderabad	12,041	1233	74.24%
Random Hindi Corpus (Siva Reddy's (IIIT hyd) Hindi Dependency Parser and POS Tagger)	1466	227	60.3%

Baseline - 50%

91.3%

84.32%-Unlabelled
78.92%-Labelled

References :

[1]Anshuman Tripathi & Manaal Faruqui : Gender prediction of Indian names. In Proceedings of the IEEE Students' Technology Symposium (TechSym) 2011, Kharagpur, India.

[2] Nivre, Joakim, Johan Hall, and Jens Nilsson. "Maltparser: A data-driven parser-generator for dependency parsing." Proceedings of LREC. Vol. 6. 2006.

[3] Husain, Samar, Prashanth Mannem, Bharat Ram Ambati, and Phani Gadde. "The ICON-2010 tools contest on Indian language dependency parsing." Proceedings of ICON-2010 Tools Contest on Indian Language Dependency Parsing, ICON 10 (2010): 1-8.

[4]Bharati, Akshar, Rajeev Sangal, Dipti Misra Sharma, and Lakshmi Bai. "Anncorra: Annotating corpora guidelines for pos and chunk annotation for Indian languages." LTRC-TR31 (2006)

[5] Begum, Rafiya, Samar Husain, Arun Dhwaj, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. "Dependency Annotation Scheme for Indian Languages." In IJCNLP, pp. 721-726. 2008.

Future Scope :

1 Sentiment Analysis Can Be Done For Determining The Gender Category Of Unrelated Nouns Existing In Any Sentence

2 Assignment Of Gender Category via Sonorants as mentioned in the last referenced paper

3 Efficient Hindi Dependency Parser

Progress Achieved -

- **Studying Basic Theory**
- **Debugging and Adding Features (Gender, Number) in Hindi Morphology Analyzer**
- **Solution to Gender of Nouns In Hindi**
- **Final Presentation And Report Submission**


THANKYOU!
Big Data
Basics
Statistics
Probability
Lifecycle
Data Science
Evolution
Types of Analytics
Role of Data Scientist
Use Cases
Mean
Median
Mode
Range
Standard Deviation
Variance
Skewness
Percentiles
Dataset
Descriptive
Inferential
Scaling
Normalization
Transformation
Outliers Detection/Fixation
Sparseness Elimination
Sampling
Exploratory Data Analysis
ANOVA
Correlation Analysis
Feature Selection
Stratified Sampling
Uncorrelated Feature Elimination
Probability Theory
Distribution
Hypothesis Testing
Significant (p-) value
Maximum Likelihood
Monte Carlo
Regression Analysis
Time Series
Continues
Discrete
Job Scheduling
Batch Processing
Model Deployment
Operationalize
Data Presentation guidelines
Visualization
Communicating Insights
Unsupervised Data Modelling
Supervised Data Modelling
Model Dev. & Validation
Causality Analysis
Correlation Analysis
Stats. & Data Exploratory Analysis
Data plotting
Descriptive Data Analysis
Unstructured Data Pre-processing
Structure Data Pre-processing
Data Pre-processing
Data acquisition and loading
Data Exploration
Frame Analytical Challenge
Mahout
Map-Reduce
MADlib
Indatabase Analytics
RHadoop
RHIPE
Hadoop Ecosystem
Zookeeper/Avro
Name & Node
Data Replication
HDFS
NoSQL
Hive/Pig
Flume
MangoDB
Cassandra
R
Data Types
Import/Export Data
Operators
Built-in Functions
Line Chart
Scatter Plot
Histogram
Boxplot
ggplot
Tree Maps
Language Detection
Keyword Extraction
Content Scraping
Concept Tagging
Social Network Analytics
Text Categorization
Sentiment Analysis
Relation Extraction
Entity Extraction & Resolution
Hidden Markov Models
Vocabulary Mapping
Tagging
Language Processing
Unstructured Data Analysis
Unstructured Data Cleansing
Corpus
Python Language
NLTK
Regular Expressions
Tuples, Dictionaries and Files
Functions
Loops
Strings and Lists
Overview
Information Retrieval
Map-Reduce
PageRank
Vector Space Model
Term Document Matrix
Term Document Frequency
Term Frequency & Weighting
Prescriptive Analytics
Mathematical Optimization
Rule-Based Expert Systems
Pattern Recognition
Supervised Feature Selection
Numerosity Reduction
Dimensionality Reduction
Machine Learning/Data Mining
Data Partition
Balance
Bias and Variance
Over-fitting
Supervised
Unsupervised
Reinforcement
Regression Analysis
Time Series
ARIMA
Exponetional Smoothing
Classifiers
K-Means Clustering
Hierarchical Clustering
Mixture Models
Artificial Neural Networks
Decision Trees & Random Forest
Bayesian Network
K-Nearest Neighbour
Support Vector Machines
Ranking
Monte Carlo
Temporal Difference
Overall Accuracy
Relative Error
Area Under the Curve
Profit
Lift
Correlation
Silhouette
Ensemble Methods
Advanced Topics
Adaptation
Meta-learning
Deep Learning