Shallow Parsing of Marathi

Harshada Gune

under the guidance of
Prof. Pushpak Bhattacharyya

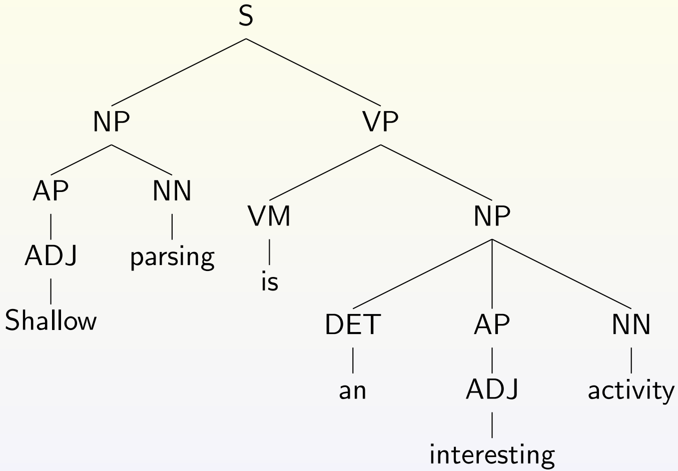
June 10, 2010

Outline

- What is Shallow Parsing?
- Literature Review
- Marathi Morphology
- Morphology Can Be Harnessed
- Architecture
- Experiments
- Conclusion
- Future work

What is Shallow Parsing?

Deep Parse


S
NP
VP
AP
NN
VM
NP
ADJ
parsing
is
Shallow
DET
AP
NN
an
ADJ
activity
interesting

Shallow Parse

[Shallow_**JJ** parsing_**NN**]_NP [is_**VM**]_VP

[an_**DT** interesting_**JJ** activity_**NN**]_NP

What is Shallow Parsing?

- Natural Language Processing (NLP) task that provides limited syntactic information
- Identifies phrases in a sentence without assigning deep hierarchical structures
- Useful and relatively tractable precursor to full parsing
- Involves two primary tasks: POS tagging and chunking

What is Shallow Parsing?

- Natural Language Processing (NLP) task that provides limited syntactic information
- Identifies phrases in a sentence without assigning deep hierarchical structures
- Useful and relatively tractable precursor to full parsing
- Involves two primary tasks: POS tagging and chunking

Example

Shallow_**JJ** parsing_**NN** is_**VM** an_**DT** interesting_**JJ** activity_**NN**.

What is Shallow Parsing?

- Natural Language Processing (NLP) task that provides limited syntactic information
- Identifies phrases in a sentence without assigning deep hierarchical structures
- Useful and relatively tractable precursor to full parsing
- Involves two primary tasks: POS tagging and chunking

Example

[**NP** Shallow parsing] [**VP** is] [**NP** an interesting activity]

What is Shallow Parsing?

- Natural Language Processing (NLP) task that provides limited syntactic information
- Identifies phrases in a sentence without assigning deep hierarchical structures
- Useful and relatively tractable precursor to full parsing
- Involves two primary tasks: POS tagging and chunking

Aim of the Work

To develop a high accuracy shallow parser for Marathi

Literature Review

Previous Work on English

(Sha and Pereira, 2003)

- Large corpora available for English
- Previous work focused on machine learning techniques
- Accuracies as high as 95-96% are obtained
- Not a morphologically rich language

Literature Review

Previous Work on English

(Sha and Pereira, 2003)

- Large corpora available for English
- Previous work focused on machine learning techniques
- Accuracies as high as 95-96% are obtained
- Not a morphologically rich language

Previous Work on Indian Languages

(Singh et al., 2006)

- Indian languages suffered from resource scarcity
- Previous work based on rule based methods
- Morphological richness challenges task of shallow parsing
- Most of the attempts used wise linguistic analysis coupled with statistical methods

General Approaches to Shallow Parsing

Statistical Methods	Rule Based Methods
(+) easy process of training	(-) rule generation is quite cumbersome process
(+) language and tag set independent	(-) dependent on language and tag set
(-) need large training data	(+) doesn't need training data
(-) not reusable to new domains	(+) usable with new domains
(-) data sparsity needs to be handled carefully	(+) no special attention is required
(-) quality of corpus matters	(-) quality of linguistic rules matters

(+) $\Rightarrow$ pros

(-) $\Rightarrow$ cons

Marathi Morphology

Marathi Morphology

Marathi Language

Highly inflectional, agglutinative, free word order

Marathi Language

Highly inflectional, agglutinative, free word order

Inflected Form	Meaning
झाडावर	on the tree
झाडाचा	of the tree
झाडाला	to the tree
झाडाने	by the tree
झाडामागे	behind the tree

Marathi Language

Highly inflectional, agglutinative, free word order

झाडासमोरच्याने = झाड + समोर + च + ने

Word	Category	Meaning
झाड	Noun	tree
झाडासमोर	Adverb	in front of the tree
झाडासमोरचा	Adjective	the one in front of the tree
झाडासमोरच्याने	Noun	by the one in front of the tree

Marathi Morphology

Marathi Language

Highly inflectional, agglutinative, free word order

Marathi Sentence	POS Sequence
मला आंबा आवडतो	PRP NN VB
मला आवडतो आंबा	PRP VB NN
आंबा मला आवडतो	NN PRP VB
आंबा आवडतो मला	NN VB PRP
आवडतो मला आंबा	VB PRP NN
आवडतो आंबा मला	VB NN PRP

Morphology Can Be Harnessed!

Morphology Can Be Harnessed!

1. Utilizing Suffixes

Suffixes contain a very good indication of the category of a word.

Morphology Can Be Harnessed!

1. Utilizing Suffixes

Suffixes contain a very good indication of the category of a word.

Motivating Examples

माणसाने उडण्याचा प्रयत्न केला.
maanasane **udnyacha_VGNN** *prayatna kela.*
man tried **flying_VGNN***.*

ण्याचा suffix identifies the correct tag

Morphology Can Be Harnessed!

1. Utilizing Suffixes

Suffixes contain a very good indication of the category of a word.

Motivating Examples

त्याने चालायला सुरुवात केली.
tyaane **chalayla_VGINF** *suruvaat keli.*
he started **to walk_VGINF**.

आयला suffix identifies the correct tag

Morphology Can Be Harnessed!

2. Restricting Categories

- POS for a word is restricted to a limited set of tags
- Morphological Analyzer (MA) produces this restricted set
- Crucial for unseen words as no explicit bias in built in model
- Classifier uses this set to narrow down the tag choice

Morphology Can Be Harnessed!

2. Restricting Categories

- POS for a word is restricted to a limited set of tags
- Morphological Analyzer (MA) produces this restricted set
- Crucial for unseen words as no explicit bias in built in model
- Classifier uses this set to narrow down the tag choice

Example

पकड - tongs - Noun
पकड - hold - Verb

Morphology Can Be Harnessed!

2. Restricting Categories

- POS for a word is restricted to a limited set of tags
- Morphological Analyzer (MA) produces this restricted set
- Crucial for unseen words as no explicit bias in built in model
- Classifier uses this set to narrow down the tag choice

Morphology Can Be Harnessed!

2. Restricting Categories

- POS for a word is restricted to a limited set of tags
- Morphological Analyzer (MA) produces this restricted set
- Crucial for unseen words as no explicit bias in built in model
- Classifier uses this set to narrow down the tag choice

Morphology Can Be Harnessed!

2. Restricting Categories

- POS for a word is restricted to a limited set of tags
- Morphological Analyzer (MA) produces this restricted set
- Crucial for unseen words as no explicit bias in built in model
- Classifier uses this set to narrow down the tag choice

Morphology Can Be Harnessed!

2. Restricting Categories

- POS for a word is restricted to a limited set of tags
- Morphological Analyzer (MA) produces this restricted set
- Crucial for unseen words as no explicit bias in built in model
- Classifier uses this set to narrow down the tag choice

Example

- Only possible categories produced by MA for **आंबा** $\Rightarrow$ {Noun}
- Hence classifier makes a confident choice even if word is unseen

Morphology Can Be Harnessed!

Ambiguity Schemes (AS)

- Set of possible categories produced by MA for a given word forms AS for that word
- Word is said to have an ambiguity when multiple POS categories possible depending upon its context
- AS for word "back" $\Rightarrow$ {Adverb,Adjective,Noun}

Morphology Can Be Harnessed!

Ambiguity Schemes (AS)

- Set of possible categories produced by MA for a given word forms AS for that word
- Word is said to have an ambiguity when multiple POS categories possible depending upon its context
- AS for word "back" $\Rightarrow$ {Adverb,Adjective,Noun}

Example

I get **back_[Adverb]** *to the* **back_[Adjective]** *seat to give rest to my* **back_[Noun]**

Architecture of Marathi Shallow Parser

Architecture

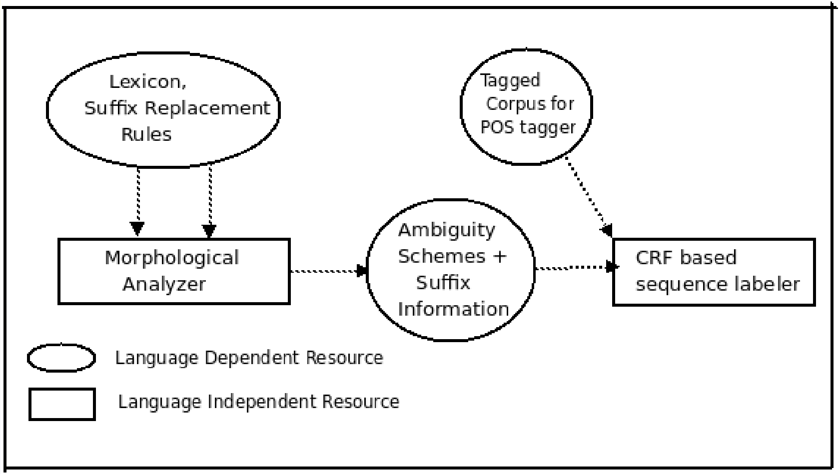
Our Methodology

- Linguistic analysis of morphosyntactic phenomena of Marathi
- Exhaustive use of morphological analyzer
- Generating rich features based on morphological analysis
- Use of POS tagged and chunk tagged data
- Training using CRF based algorithm

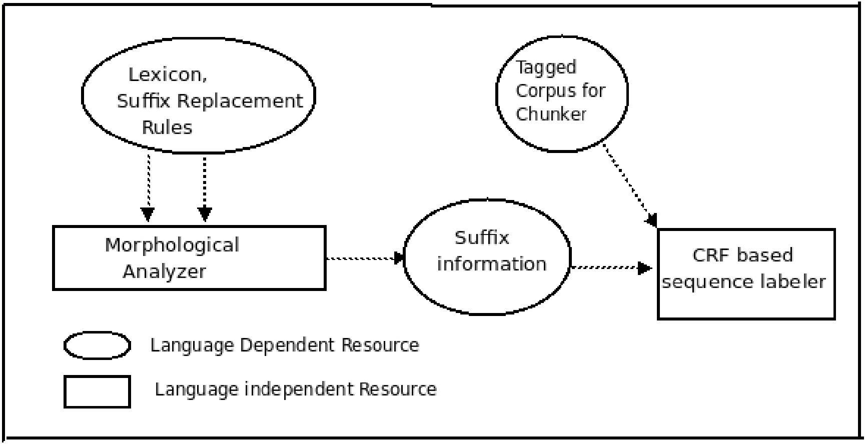
Resources Used

- **Lexicon:**
  Stores root words, their paradigm and and category information
- **Suffix Replacement Rules:**
  Encodes the information needed to get the root from inflected word
- **Training Data:**
  POS and chunk tagged data

POS Tagger


Lexicon,
Suffix Replacement
Rules
Tagged
Corpus for
POS tagger
Morphological
Analyzer
Ambiguity
Schemes +
Suffix
Information
CRF based
sequence labeler
Language Dependent Resource
Language Independent Resource

Chunker



Lexicon,
Suffix Replacement
Rules
Tagged
Corpus for
Chunker
Morphological
Analyzer
Suffix
information
CRF based
sequence labeler
Language Dependent Resource
Language independent Resource

Experiments: POS Tagging

Experiments: POS Tagging

Features used for POS tagger

- $t_i\ t_{i-1}$ and w_j such that $i-2 < j < i+2$
- $t_i\ t_{i-1}$ and suffix information of w_i
- $t_i\ t_{i-1}$ and ambiguity scheme of w_i

Experiments: POS Tagging

Feature Variations for POS Tagging

Weak Features (WF)

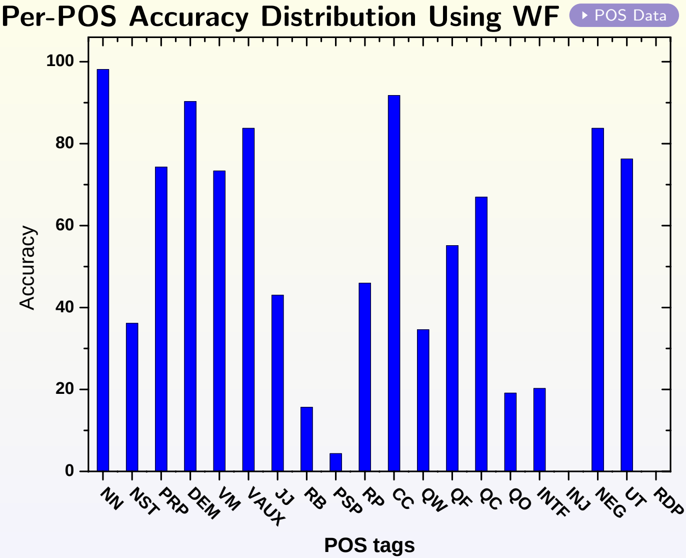
- Word and bi-gram tag features
- Overall accuracy = 85%

Morphological Features (MF)

- Suffix information and Ambiguity Schemes (AS) added to WF
- Overall accuracy = 95%

POS Data

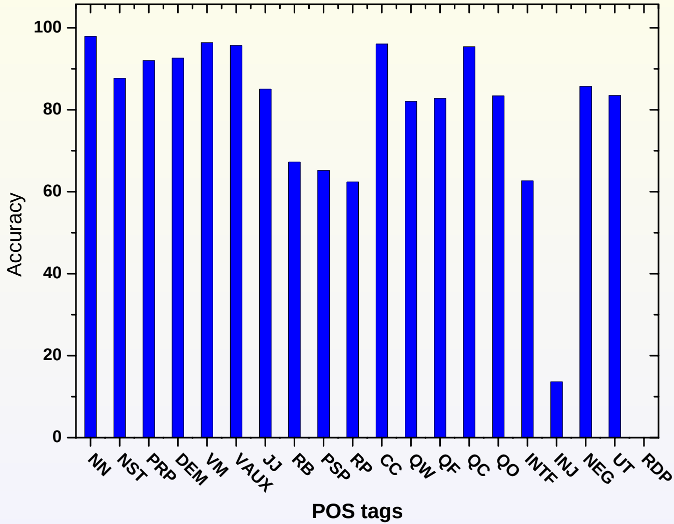
Experiments: POS Tagging



Experiments: POS Tagging

Per-POS Accuracy Distribution Using MF

POS Data



POS Data

	NN	NST	PRP	DEM	VM	VAUX
NN	50092	0	63	1	621	23
NST	337	209	8	0	21	0
PRP	1756	0	6515	341	68	16
DEM	99	0	207	2926	4	1
VM	3876	0	3	8	12995	807
VAUX	271	0	1	1	748	5273

Table: POS Confusion Matrix with WF

Experiments: POS Tagging

POS Data

	NN	NST	PRP	DEM	VM	VAUX
NN	49988	18	92	2	167	4
NST	33	507	9	0	3	0
PRP	145	3	8071	312	8	5
DEM	3	0	231	3002	2	1
VM	225	1	4	9	17078	347
VAUX	10	0	1	1	257	6025

Table: POS Confusion Matrix with MF

Experiments: POS Tagging

▸ POS Data

POS Tag	Errors in unseen words [1] (in %)
NST	100
PRP	100
VM	63
VAUX	77
JJ	98
RB	100
QW	100
QF	100

Table: Unseen Words Statistics with WF

[1] Words not present in training data

Experiments: POS Tagging

POS Data

POS Tag	Errors in unseen words [1] (in %)	
	WF	**MF**
NST	100	52
PRP	100	32
VM	63	8
VAUX	77	31
JJ	98	38
RB	100	61
QW	100	46
QF	100	67

Table: Unseen Words Statistics with WF and MF

[1] Words not present in training data

Experiments: Chunking

Experiments: Chunking

Features used for Chunker

- c_i c_{i-1} and w_j such that $i - 1 < j < i + 1$
- c_i c_{i-1} and t_j such that $i - 1 < j < i + 1$
- c_i c_{i-1} and suffix information of w_i

Feature Variations for Chunking

Weak Features (WF)

- Word, POS and bi-gram tag features
- Overall accuracy = 96.91%

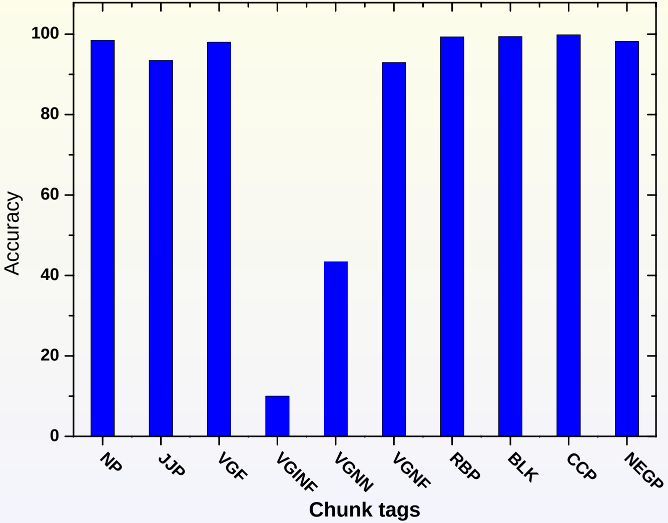
Morphological Features (MF)

- Suffix information added to WF
- Overall accuracy = 97.8%

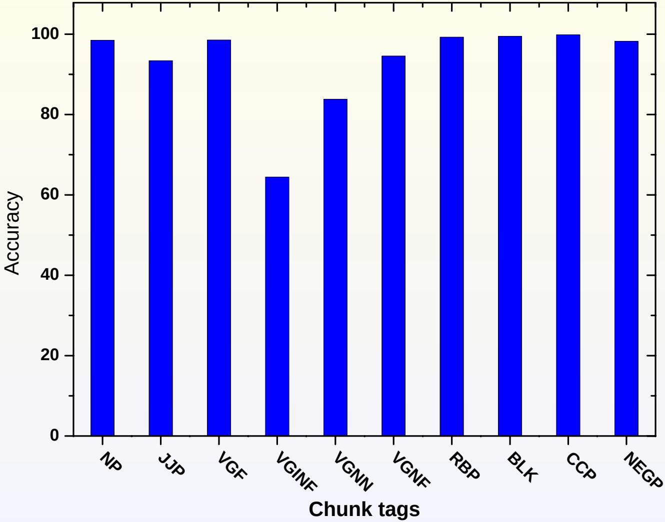
Chunk Data

Per-Chunk Accuracy Distribution Using WF

▸ Chunk Data



Per-Chunk Accuracy Distribution Using MF Chunk Data



Experiments: Chunking

Chunk Data

	VGF	VGINF	VGNN	VGNF
VGF	20783	0	23	242
VGINF	13	9	16	59
VGNN	280	0	797	850
VGNF	350	5	99	5241

Table: Confusion Matrix for Chunking with WF

Chunk Data

	VGF	VGINF	VGNN	VGNF
VGF	20857	0	39	150
VGINF	11	58	7	21
VGNN	163	0	1570	194
VGNF	229	14	106	5347

Table: Confusion Matrix for Chunking with MF

Experiments: Chunking

Using only POS information

Features

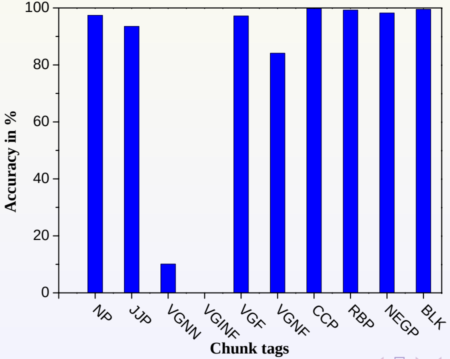
- c_i and t_j such that $i - 1 < j < i + 1$
- c_i c_{i-1} and t_j such that $i - 1 < j < i + 1$
- Chunkwise accuracy $= 95\%$

Experiments: Chunking

Using only POS information

Features

- c_i and t_j such that $i-1 < j < i+1$
- c_i c_{i-1} and t_j such that $i-1 < j < i+1$
- Chunkwise accuracy $= 95\%$



Experiments: Chunking

Using only POS information

Features

- c_i and t_j such that $i - 1 < j < i + 1$
- c_i c_{i-1} and t_j such that $i - 1 < j < i + 1$
- Chunkwise accuracy $= 95\%$

Pointer to language adaptation work!

Experiments: Linguistic Analysis vs Data Size

Linguistic Wisdom vs Statistical Brawn

Linguistic knowledge obviates large size corpora

Use of suffix information and Ambiguity Scheme

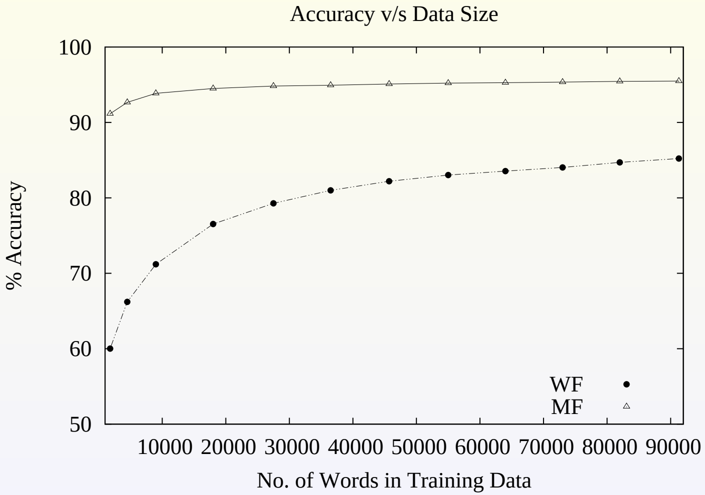
POS Tagging

- Accuracy of only 85% obtained with WF using around 91k words
- Accuracy as high as 94% obtained with MF using only 20k words

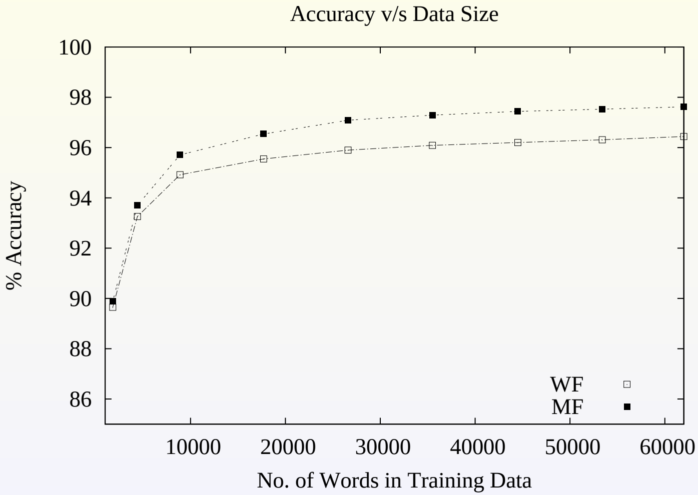
Chunking

- 60k words data needed to get 96% accuracy with WF
- Same accuracy is achieved using only 20k words with MF

Average Accuracy of all POS Tags



Average Accuracy of all Chunk Tags



Importance of verbs

Verb POS Tags

- Accuracy of only 79% obtained using 91k words with WF
- Accuracy of around 95% is obtained using only 10k words with MF

Verb Chunks

- Around 60k words needed to get accuracy of 90% with WF
- Same accuracy is achieved using only 10k words with MF

Linguistic Wisdom vs Statistical Brawn

Importance of verbs

Verb POS Tags

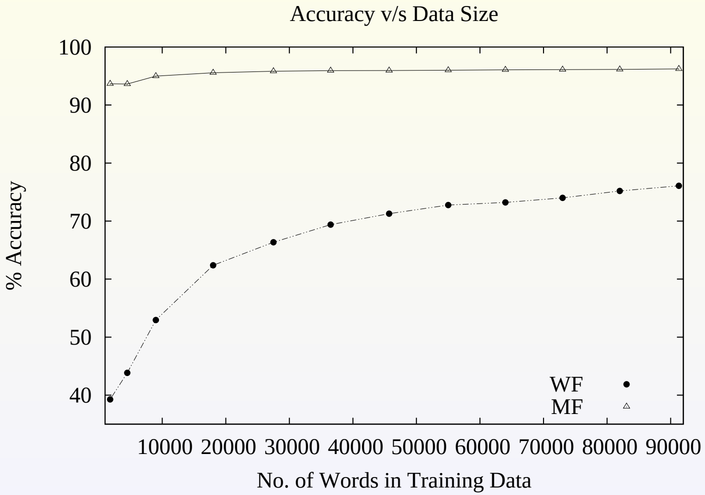
- Accuracy of only 79% obtained using 91k words with WF
- Accuracy of around 95% is obtained using only 10k words with MF

Verb Chunks

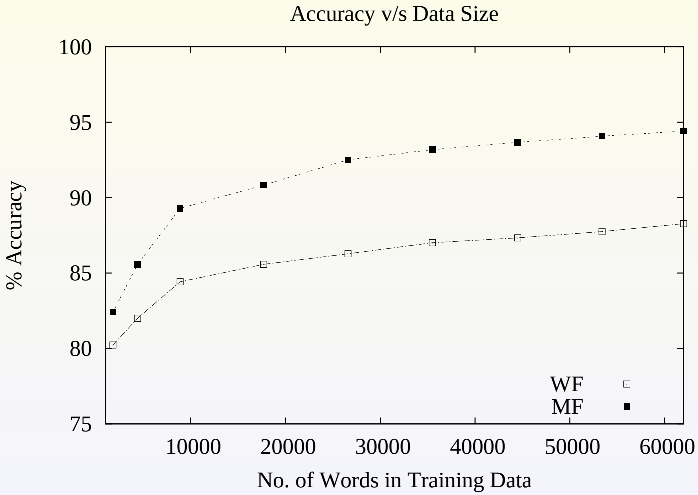
- Around 60k words needed to get accuracy of 90% with WF
- Same accuracy is achieved using only 10k words with MF

Verbs are where all the action lies!

Average Accuracy of Verb POS Tags



Average Accuracy of Verb Chunk Tags



Conclusion

- Shallow parsing provides the partial syntactic information about the sentence
- Useful in information extraction, information retrieval, named entity recognition, machine translation etc.
- Morphological richness of Marathi imposes some challenges in building high accuracy shallow parser
- The task becomes easier if features of language are harnessed properly
- For morphologically rich languages linguistic wisdom can overpower statistical brawn
- POS tagger with accuracy of 95% and chunker with accuracy of 98% are built for Marathi

Future Work

- Further scope of improvement in noun group of POS tagging
- Experiments in chunking with only POS information can be extended to language adaptation work: useful in resource poor scenario
- "Linguistic analysis vs data size" can be tested on other Indian languages

This work has been accepted in
Computational Linguistics Conference (COLING 2010), Beijing, China, August 2010

Verbs are where all the Action Lies: Experiences of Shallow Parsing of a Morphologically Rich Language

- Harshada Gune, Mugdha Bapat, Mitesh Khapra and Pushpak Bhattacharyya,

Thank You!

References

- Bharati, A., Chaitanya, V., and Sangal, R. (1995). Natural Language Processing : A Paninian Perspective. Prentice Hall India.
- Navanath, S., Dhrubajyoti, D., Utpal, S., and Jugal, K. (2009). Part of Speech Tagger for Assamese Text. In Proceedings of the ACLIJCNLP 2009 Conference Short Papers, pages 33-36, Suntec, Singapore. Association for Computational Linguistics.
- Berwick, I. R., Abney, S., (eds, C. T., and Abney, S. P. (1991). Parsing By Chunks.
- Damale, M. K. (1970). Shastriya Marathi Vyaakarana. Pune Deshmukh and Company.

- Dandapat, S., Sarkar, S., and Basu, A. (2007). Auto- matic Part-of-Speech Tagging for Bengali: An Approach for Morphologically Rich Languages in a Poor Resource Scenario. In ACL. The Association for Computer Linguistics.
- Dixit, V., Dethe, S., and Joshi, R. K. (2006). Design and Implementation of a Morphology-based Spellchecker for Marathi, an Indian Language. In Special issue on Human Language Technologies as a challenge for Computer Science and Linguistics. Part I. 15, pages 309-316. Archives of Control Sciences.
- Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proc. 18th International Conf. on Machine Learning, pages 282-289. Morgan Kaufmann, San Francisco, CA.
- Pawar, S. S. (2008). Part of Speech Tagging of Indian Languages. Master's thesis, Computer Science Department, IIT Bombay.

References

- Sha, F. and Pereira, F. C. N. (2003). Shallow Parsing with Conditional Random Fields. In HLT-NAACL.
- Shrivastava, M. and Bhattacharya, P. (2008). Hindi POS Tagger Using Naive Stemming: Harnessing Morphological In- formation Without Extensive Linguistic Knowledge. In Proceedings of the ICON.
- Singh, A., Bendre, S., and Sangal, R. (2005). HMM Based Chunker for Hindi. In Proceedings of International Joint Conference on NLP.
- Singh, S., Gupta, K., Shrivastava, M., and Bhattacharyya, P. (2006). Morphological Richness Offsets Resource Demand - Experiences in Con- structing a POS Tagger for Hindi. In Proceedings of ACL-2006
- SPSAL (2007). Workshop On Shallow Parsing for South Asian Languages (SPSAL).
- Valambe, M. R. (2007). Sugam Marathi Vyaakarana Lekhan. Nitin Prakashan, Pune.

Data Set

POS Tag	Frequency in Corpus	POS Tag	Frequency in Corpus
NN	51047	RP	359
NST	578	CC	3735
PRP	8770	QW	630
DEM	3241	QF	1928
VM	17716	QC	2787
VAUX	6295	QO	277
JJ	7311	INTF	158
RB	1060	INJ	22
UT	97	RDP	39
PSP	69	NEG	154

Table: POS Tags in Training Data

Back

Data Set

Chunk Tag	Frequency in Corpus	Chunk Tag	Frequency in Corpus
NP	40254	JJP	2680
VGF	7425	VGNF	3553
VGNN	1105	VGINF	58
RBP	782	BLK	2337
CCP	4796	NEGP	43

Table: Chunk Tags in Training Data

Back

Chunk Examples

Chunk Type	Tag Name	Example
Noun Chunk	NP	(हे_DEM **नवीन** _JJ **पुस्तक** _NN)_NP
Adjectival Chunk	JJP	दिवस_NN (**मस्त** _JJ)_JJP गेला_VM
Finite Verb Chunk	VGF	मी_PRP घरी_NN (**जेवले**_VM)_VGF
Non Finite Verb Chunk	VGNF	तो_PRP (**खेळताना**_VM)_VGNF पडला_VM
Infinitival Verb Chunk	VGINF	मला_PRP (**गायला**_VM)_VGINF आवडते_VM
Gerund Verb Chunk	VGNN	(**लिहायच्या**_VM)_VGNN त्रासातून_NN सुटका_NN
Adverb Chunk	RBP	तो_PRP (**हळूहळू**_RB)_RBP चालतो_VM
Conjunct Chunk	CCP	राम_NNP (**आणि** _CC)_CCP श्याम_NNP ख़ेळतात_VM
Miscellaneous	BLK	नदी_NN (**जणू**_UT)_BLK आमची_PRP आईच_NN

Applications of Shallow Parsing

- **Summary Generation and Question Answering:** information about specific syntactic-semantic relations such as agent, object, location, time etc. is required
- **Named Entity Recognition:** as a preliminary to NER to pick out noun phrases from a text
- **Machine Translation:** identifying the specific constituents in the sentence that has to undergo transformation.