Chunking Marathi Text using Marathi Grammar Rules and Conceptual Dependency Representation

Madhuri M. Deshpande, Sharad D. Gore

Abstract- The paper aims at using a rule based chunker to create chunks, such as noun phrase (NP) and verb phrase (VP), for a given Marathi sentence. Chunking is the process that labels segments of a sentence with syntactic constituents such as noun/verb/adjective phrases (NP, VP, AP). Chunking is an important task in Natural Language Processing (NLP). We have used a modified YASS POS tagger to tag the input Marathi sentence. We have applied Marathi grammar rules, in the form of regular expressions, and used the Conceptual Dependency (CD) theory representation to represent the dependency of words in Marathi sentence, which ultimately would depict the meaning of words in the sentence with respect to the context in which a bag-of-words are used. A rule-based chunker create chunks in Marathi sentence. Conceptual Dependency Theory focuses on concepts and understanding about a concept instead of syntax and structure.

***Index Terms*: Chunking, Conceptual Dependency, Dependency Parser, Natural Language Processing.**

I. INTRODUCTION

Chunking, a kind of shallow syntactic analysis, aims at labelling segments of a sentence with syntactic constituents such as noun/verb/adjective phrases (NP, VP, AP). These segments are sometimes referred to as groups. Chunking (or local word grouping) is also called partial parsing, It reduces the computational effort at the level of parsing by assigning partial structure to a sentence. A typical chunk, as defined by Abney (1994:257) *consists of a single content word surrounded by a constellation of function words, matching a fixed template.* Chunks are considered as the truncated versions of typical phrase-structure grammar phrases that do not include arguments or adjuncts (Grover and Tobin 2006). Chunks can be represented as connected sub-graphs of a sentence's parse tree. They are defined in terms of major heads and have their own syntactic structure that can be represented in the form of a tree. Each word in a sentence is assigned only one unique tag, often encoded as a begin-chunk (e.g., B-NP) or inside-chunk tag (e.g., I-NP). Two heads of the same lexical category are not allowed inside a chunk. These groups are non-overlapping, i.e., a word can only be a member of one chunk (Sang and Buchholz, 2000). The chunk structures are also flat and non-recursive. Not all words in a sentence need belong to a group. Some words in a sentence may not be classified into any of the identified chunks.

Revised Version Manuscript Received on November 07, 2016.

Mrs. Madhuri M. Deshpande, Department of Computer Science, Savitribai Phule Pune University (formerly University of Pune), Pune, India.

Dr. Sharad D. Gore, Department of Computer Science, Savitribai Phule Pune University (formerly University of Pune), Pune, India.

A chunker uses the POS information provided by a tagger to form groups. A Noun Group consists of a head noun along with its qualifiers and modifiers (including particles). A Verb Group contains a single main verb and any auxiliaries, negation markers, and focus particles. The grammatical information of a word group depends on the order of its constituent morphemes and the information associated with those constituents. The group identification module makes use of the morphological information and the POS information provided by a morphological analyser and a POS tagger respectively.

Chunking provides a key feature that helps on more elaborated NLP tasks such as parsing and information extraction. It is closely related to the task of Named Entity Recognition (NER) and can contribute to improving the performance of NER taggers (Zhou and Su, 2002).

Chunking systems may be rule-based (Abney, 1996; Finch and Mikheev, 1997) or based on machine learning. Machine learning chunkers may be symbolic (D´ejean, 2000; Johansson, 2000; Vilain and Day, 2000) or statistical (the majority of recent systems). Bharati et al. (1995) designed the first Local word grouping in Hindi. They used a morphological analyser and a Local word grouper (LWG) to build the Paninian Parser. A list of regular expressions was used to form local word groups by Ray et al. (2003). They worked on the five structures that rely only on local modifier-modified relationships and do not need long distance dependencies.

Statistical models have also been used in Hindi word grouping has also been attempted by Baskaran (2006), that used the HMM based approach. Singh A. et al. (2005) and Dalal et al. (2006) used the Maximum Entropy Models as a local word grouper. These machine learning systems use a very large corpus and very little linguistic knowledge. CRF based POS tagging system by Gune H. et. al. (2010) was developed for Marathi verb group identification. Limited noun phrase chunking has also been done for Tamil (Vijay and Sobha 2010).

In this paper we propose to use the Conceptual Dependency theory representation of NLP sentences coupled with Marathi grammar rules to create relevant chunks for Marathi sentences. We have used a modified YASS POS tagger to tag the input Marathi text. This paper defines hand-crafted rules for identifying Noun Groups and Verb Groups in a Marathi sentence with the help of morphological rules and Marathi grammar rules that apply in a particular

context. This experiment will help in resolving the ambiguities arising in POS tags. We are using a range of well-understood grammatical features of the Marathi Language.

II. A BRIEF REVIEW OF CD THEORY

Schank's (1975) Conceptual Dependency (CD) theory[14] was developed as part of a natural language comprehension project. Conceptual dependency theory allows one to represent the meaning of natural language sentences in a way that facilitates drawing inferences from the sentences. That is, CD representations ought to explicitly represent implicit information that will be used in drawing inferences in order to produce complete representations. If two sentences have the same meaning, they should be represented the same, regardless of the particular words used. CD representation is independent of the language in which the sentences were originally stated.

Most parsers used in "conversation" machines have been syntactic parsers. That is, they analyze an input sentence and construct a syntactic (grammatical) network from it. But consider the sentence, "I hit girl with long hair with a hammer with vengeance." Clearly the syntactic structure of this sentence will not provide the information necessary to get at the meaning of it. For example, if we need to know that it was the hammer that hit the girl, we would have to use methods more sophisticated than syntactic analysis. Whatever the purpose, it is the *meaning* of the input sentence that is needed, not its syntactic structure.

On the sentential level, the sentences of a given language are encoded within the syntactic structure of that language. The basic construction of the sentential level is the sentence. The next higher level in the system is the conceptual level, and the basic construction of this level is the conceptualization. A conceptualization consists of concepts and certain relations that exist between these concepts. Underlying every sentence in a language there exists at least one conceptualization.

The conceptual level works with a system of rules that operate on conceptual categories. These rules generate all the permissible dependencies in a conceptualization[14]. Multiple combination of conceptualizations in various relationships are intended to account for the totality of human language activity at the conceptual level. The vocabulary for conceptual dependency consist of the following:

- a set of *primitives,* shown in table 1, which are used to represent actions in the world
- a set of *states* used to represent preconditions and results of actions
- a set of *dependencies* or possible conceptual relationships which could exist between primitives, states and the objects involved.

CD Theory uses four **primitive conceptualizations viz.** actions (**ACT**: actions), objects (**PP**: picture producers), modifiers of actions (**AA**: action aiders) and modifiers of objects (**PA:** picture aiders).

Representations of sentences could be constructed by piecing together these building blocks to form a *conceptual dependency graph.*

Table 1. The CD primitives [14]

Primitive	Meaning
PTRANS	The transfer of location of an object
ATRANS	The transfer of ownership, possession, or control of an object
MTRANS	The transfer of mental information between agents eg. read, see
MBUILD	The construction of a thought or of new information by an agent eg. think, assume
ATTEND	The act of focusing attention of a sense organ toward an object
GRASP	The grasping of an object by an actor so that it may be manipulated
PROPEL	The application of a physical force to an object
MOVE	The movement of a body part of an agent by that agent
INGEST	The taking in of an object (food, air, water, etc.) by an animal
EXPEL	The expulsion of an object by an animal
SPEAK	The act of producing sound, including non-communicative sounds

Each primitive had a set of *slots* associated with it, from the set of conceptual dependencies. Associated with each slot are restrictions as to what sorts of objects could appear in that slot. For example, the slots for PTRANS are:

- ACTOR: a HUMAN (or animate object), that initiates the PTRANS
- OBJECT: a PHYSICAL OBJECT, that is PTRANSed (moved)
- FROM: a LOCATION, at which the PTRANS begins
- TO: a LOCATION, at which the PTRANS ends.

Conceptual dependency representations are written graphically as shown in Figure 1. The actor of a primitive action was connected to the primitive using a double arrow; the object appeared to the right with a single arrow connection, and the source and destination (TO and FROM) appeared to the right of the object.

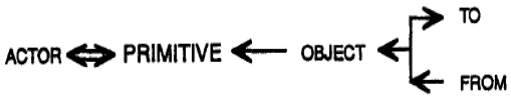


Figure 1: Basic form of CD Graph (Schank)

Objective dependency is denoted as ←, attributive dependency and is denoted by ↑, Prepositional dependency between two PP's is denoted as ⇑ with a label indicating the type of prepositional dependency.

The sentence, "वरदने गगनला पुस्तक दिले होते . ते त्याने हरवले हो ते " would be represented as shown in Figure 2. वरद was explicitly represented as both the ACTOR and FROM slots as

वरद had possession of the book originally.

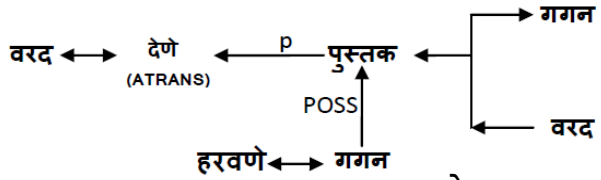


Figure 2: CD Representation of "वरदने गगनला पुस्तक दिले होते . त्याने ते हरवले होते "

Figure 3 shows the CD representation of रामाने रावणाचा वध केला व त्याने अयोध्येस प्रस्थान केले . From figure 3, it becomes clear that the 'त्याने ' refers to 'राम ' instead of 'रावण ' since 'रावण ' is physically in a dead state and cannot travel to 'अयोध्या '.

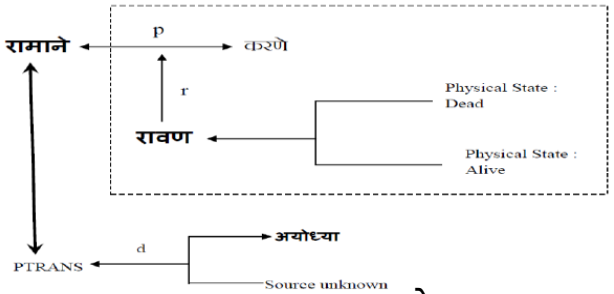


Figure 3: CD representation of रामाने रावणाचा वधकेला व त्याने अयोध्येस प्रस्थान केले [7].

The conceptual categories are divided into governing and assisting groups:

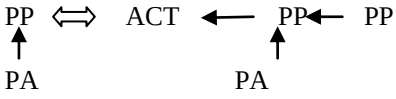
Governing Categories

PP	An actor or object; corresponds syntactically (in English) to concrete nominal nouns or noun forms.
ACT	An action; corresponds syntactically (in English) to verbs, verbal nouns, and most abstract nouns.
LOC	A location of a conceptualization
T	A time of conceptualization; often has variant forms consisting of parts of a conceptualization

Assisting Categories

PA	Attribute of a PP; corresponds (in English) to adjectives and some abstract nouns.
AA	Attribute of an ACT; corresponds (in English) to adverbs and indirectly objective abstract nouns.

Thus, the categories assigned in the above network correspond closely to their syntactic correlates:



A conceptualization is written in a conceptual dependency analysis on a straight line. Dependents written perpendicular to the line are attributes of their governor except when they are part of another conceptualization line. Whole conceptualizations can relate to other conceptualizations as actors . The representation is in terms of actor-action-object conceptualizations in a topic-cogent form. Thus, words that have many syntactic forms will have only one conceptual form. This is true inter-linguistically as well as intra-linguistically.

The conceptual level is intended to represent the concepts and relations between concepts that underlie natural language sentences. There are formally defined dependency relations between given categories of concepts, and these relations are the conceptual rules [7].

III. ENTITY-ATTRIBUTE-VALUE STRUCTURE

All information represented in Fig. 3 is held in "records", which are lists of attribute-value pairs. Some records represent physical entities, some abstract entities such as actions. For every word there is a structure and values of attributes of the record specify characteristics of the word.

There is a record for each concept in the structure. A SUP erset attribute relates concepts to other concepts. A record may also describe an entire conceptualization.

For example, the sentence "वरदने गगनला पुस्तक दिले होते " could be represented by the record:

(SUP देणे , AGENT वरद , GOAL पुस्तक , RECIPIENT गगन, PAST)

Here, PAST is an indicator that indicates existence of a particular condition.

Other attributes that may hold in conceptualization are LOCATION, TIME, SOURCE, DESTINATION, DONOR, REASON, INSTRUMENT, CONCURENT and many others.

This process of converting Marathi sentence into a record structure representing the meaning of the input is called "decoding".

In the sentence 'त्या ने ते हरवले होते ', the 'त्याने ' refers to ' गगन ' and not 'वरद ' since the CD structure of the previous sentence is showing the book to be in possession of 'गगन '. The concept represented in this sentence is 'हरवणे ' and the object is 'पुस्तक '. These inferences can be made by just referring to the slot structure of the previous sentence. Now, the sentence becomes 'गगनने पुस्तक हरवले होते '. This we call as paraphrasing. The slot representation becomes (SUP हरवणे , AGENT गगन, GOAL पुस्तक , RECIPIENT unknown, PAST).

IV. THE INFORMATION STRUCTURE

A conceptualization is represented in the Internal Data Structure (IDS) of NLP as a set of records representing the concepts of that conceptualization. The action is considered to be the most important building block in the structure. The IDS representation of the sentence 'वरदने गगनला पुस्तक दिले होते ' is as follows:

R1('देणे ', AGENT='वरद', GOAL='R2', RECIPIENT='गगन', PAST) R2('पुस्तक ')

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Here, R1 & R2 represent the concepts involved. Consequently, there is a special record ('ACTNLIST') in which there are attributes pointing to each action record. Each of these action records has attributes linking the action to the other concepts in the conceptualization. Figure 4 shows graphically the IDS representation for the conceptualization 'वरदने गगनला पुस्तक दिले होते ' and the records Rl and R2 that would be created by NLP during the decoding of the sentence. Rl would be pointed to by 'ACTNLIST'. Here we represent each relation in a conceptualization of CD theory by an NLP record with specified attributes.

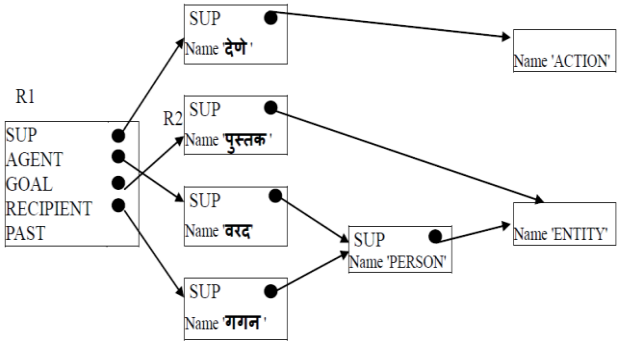


Figure 4: IDS representation of 'वरदने गगनला पुस्तक दिले होते'

V. THE CONCEPTUAL DEPENDENCY PARSER FOR MARATHI TEXT

The Conceptual Dependency parser is a conceptual analyser and a syntactic parser. It extracts the underlying meaning and conceptual relationships present in Marathi text. The parser's output is a language-free network consisting of unambiguous concepts and their relations to other concepts. Marathi sentences having identical meanings parse into the same conceptual network.

The CD theory framework can act as a generative theory. To do this we need to add semantics and realization rules. The realization rules are used in conjunction with a dictionary of realizates. These rules map pieces of the network in accord with the grammar[18]. Thus, a simple rule in English might be:

PP
↑ = Adj + N
PA

By reversing the realization rules and using the semantics as a check with reality, the Conceptual Dependency framework is used as a natural language parser. The system for analysing a Marathi sentence into its conceptual representation operates on pieces of a sentence looking up the potential conceptual realizates.

A realization rule consists of two parts: a recognizer and a dependency chart. The recognizer determines whether the rule applies and the dependency chart shows the dependencies that exist when it does. In the recognizer, we specify the ordering, categories, and inflection of the concepts and connectives that normally would appear in a sentence if the rule applied. If certain concepts or connectives are can be omitted in the input, the rule can specify what to assume when such connectives are missing.

Realization rules are used to fit concepts into the network and to anticipate further concepts and their potential realizates in the network. Concepts are fitted into the network as and when they are encountered. When a rule is selected for the current word sense, it is compared with the rules of preceding word senses to find one that "fits". For example, if "very good" is seen, one realization rule for "very" is:

very PA = 1
↑
0

The tags 0 and 1 indicate relative order of word sense and identify them for reference in the dependency chart. A negative order number for a word indicates preceeding word(s). 0 means current word. One rule for 'hot' is :

AA PA = 0
-1 0 ↑
-1

VI. IDENTIFYING AMBIGUITY

The method for group level analysis of Marathi text to resolve the ambiguities between adjective and noun, main verb and auxiliary verb or demonstrative and pronoun is described here.

VII. NOUN GROUPS (NGS) IN MARATHI

A. Demonstrative-Personal Pronoun POS ambiguity:

In the sentence, 'त्याला वाचवणे आवश्यक होते ' the tagger tags the word 'त्याला ' properly as PRON (pronoun). However, for a sentence begining with a demonstrative such as 'त्या सुंदर चित्रातील .... ', 'त्या कोपर्यातील बाकावर .... ' the tagger tags the word 'त्या ' also as PRON instead of DEM (demonstrative).

B. *Adjective-Noun ambiguity:*

It was observed that sometimes an adjective may function as a head noun if the noun is dropped, and bears the same inflection as the nominal head. For example, in the sentence 'चांगल्या कामाचा परिणाम चांगलाच असतो ' the word 'चांगल्या ' is treated as a nominal head instead of an adjective.

C. Noun-Verb ambiguity:

Many nouns may appear as verbs even when inflected. Consider the following sentences:

i) माझे महाराष्ट्र बँकेत खाते होते .

The word 'खाते ' appears as a VERB (habitual, plural) instead of NOUN (plural, direct).

ii) १७ तासांच्या कठोर परिश्रमानंतर , मला झोप आवश्यक होती .

The word 'झोप ' is correctly tagged as VERB and is part of verb group. However, in the following sentence the same word (झोप) appears as verbal NOUN instead of VERB. But the tagger tags the verbal noun as verb.

शरीराला कमीतकमी ६ तास झोप आवश्यक आहे .

To conclude, verbs may appear as verbal nouns in their infinitival form and may function as nouns. Nevertheless, a verbal noun retains many of its verbal properties. In the sentence 'पोह णायाचे शरीराला अनेक फायदे आहेत ' the word 'पोह णे ', appearing as 'singular /oblique' case, is tagged as a VERB instead of NOUN.

VIII. SOME RULES FOR IDENTIFYING NOUN GROUPS (NGS)

Noun groups typically have a main word which can be a noun or pronoun around which other words in the noun group relate to. Noun groups may have either of the following structure:

- N e.g. पुस्तक , कपाट , ... (PP)
- Dem + Pron+ N e.g त्या पुस्तकात . . . (PP + PP)
- Poss Pronoun + N e.g. माझ्या घरीं (PP + PP)
- Adj + N e.g. कठोर निर्णय (PA+PP)
- Dem Pron + Adj + N e.g. तो कठोर निर्णय (PP + PA + PP)
- Card + N e.g. दोन खोल्या (PA +PP)
- Ord + Noun e.g. पहिले बक्षीस , दुसर्या स्थानी (PA + PP)
- Non-Spec Det + N e.g. काही निर्णय , अनेक वेळा (PA + PP)
- Det + Adj + N e.g. काही घाण सवयी , काही चांगल्या सवयी , काही जुन्या आठवणी
- Inten + Adj + N e.g. खुप जुन्या आठवणी
- Pron e.g. त्या , ती , ते , तो .....
- N or Proper N e.g. मुलगा , हरी etc.

The ordering of the constituent elements of a Marathi NG can be captured using the dependency chart processed by the dependency parser of the CD representation.

We place candidate constituents of the Marathi NG in four sets as shown below. Parenthesis mark optional elements. Two or more elements always appearing together are shown as a single unit within parentheses.

Set 1 includes possessive demonstrative pronouns. Both are optional elements of a Marathi NG.

Set 2 includes numerals and intensifiers. A numeral word is usually an approximate, ordinal, cardinal fractional, quantifiers, multiplicative, aggregative or a measure word.

Set 3 includes adjectives (including imperfective or perfective verbal adjectives) and are optional in an NG and may recursively appear inside an NG.

Set 4 includes nouns and pronouns which are obligatory members of an NG.

We now apply the following regular expression to determine the word grouping rules for a Marathi NG.

* stands for zero or more repetitions.

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

Here, we assume that every Marathi NG is a non-recursive group with its specifiers and modifiers and only one head.

We can now write an algorithm using the above findings to identify noun groups in a Marathi sentence. The input to this algorithm will be output of the Part-of-speech (POS) tagger and the IDS representation for a sentence. For each word, the POS tagger gives the stem and the set of suffixes along with the associated morphological properties. The output will be a nearly exact NG(s) for an input Marathi sentence.

IX. ALGORITHM FOR NG IDENTIFICATION

1. For a tagged Marathi sentence, start form the extreme right of the sentence and look for :
   - A Set 4 element to start an NG
     If Set 4 element is found, a NG has started
   - If a Demonstrative pronoun is found
     Consider it as a Pronoun (head)
2. If a NG has started
   - Look for a Set 3 and/or Set 2 and/or Set 1 element
   - If Set 3, 2 and 1 elements are found then
     The NG includes the current word
   - If set 3, 2 and/or 1 elements are not found then
     The NG has already ended with the previous word
3. If a NG is completely identified
   Check with the network constructed by the dependency parser. In case of a disagreement, resolve the disagreement using the IDS structure.
   Apply the regular expression to cross check the NG.
5. Start looking for the next NG.
6. Process stops when all NGs in the Marathi sentence are identified.

X. VERB GROUPS (VGS) IN MARATHI

A Marathi VG includes a single main verb root followed by a sequence of inflectional suffixes and/or auxiliary verb sequences. The group contains various verbal morphemes that centre on a single event and the order of occurrence of these verbal morphemes is fixed. The verbal morphemes are subject to several grammatical and semantic constraints. Some examples of Marathi VGs are पुस्तक ठेवले (kept the book) (ठेवणे -past) and ठेवले गेले होते (ठेवणे -perfect passive-perfect, singular).While analysing Marathi VGs, we have considered simple predicates and for the persent study we have ignored conjunct verbs or compound verbs.

XI. IDENTIFYING VERB GROUP (VG) BOUNDARIES

A VG boundary is marked by analysing how the verbs order themselves in Marathi sentence by examining the tense, aspect, mood, modality, gender, number, person, voice and finiteness. The linear order of the major grammatical categories within a Marathi VG is *Verb-Aspect-Tense/Mood.* A Marathi verb group must always begin with a main/root verb with or without a suffix. Once the main verb is identified, the verb group begins. Scanning from left to right, the main verb may be followed by a string of

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

intermediate verbal suffixes and follow a linear order.

The root verb, inflected or not, forms the 'start' marker for a Marathi VG. All verbal auxiliaries may also be considered as start markers. Since the identification begins from left to right, the first instance of a free verbal morpheme is always the root or main verb as shown in the following example.

जगनघोड्याला [मारत होता].

Here, we ignore poetic constructions, sarcastic sentences and proverbs.

A. Intermediate markers:

These markers include possible-end markers and 'must continue' markers. Possible end markers are those which may end a VG such as the *perfective* marker or the modal auxiliary for *necessity* (preceded by an infinitive-gender, number sequence). These intermediate markers may be followed by other intermediate markers to further extend the VG. For example, the *perfective* marker may be followed by the past or the present tense auxiliary as in तो जेवला (he ate), तो जेवत आहे (he is eating)and तो जेवला होता (he had eaten). A past tense auxiliary can precede a modal auxillary, such as त्याने काम करायला हवे होते (he should have done the work) and the subjunctive marker may be followed a *future-person, number* marker, such as जेवेल , बघेल , वाचेल etc. The 'must-continue' markers must be followed by other verbal morphemes in order to complete the VG.

The verbal elements appear in a specific order. This ordering is subject to a number of constraints as listed below:

B. Specific Constraints within a Marathi VG

a) The modal auxiliary 'पाहिजे ' must be preceded by an infinitive (with gender-number) marker, such जेवले पाहिजे / झोपले पाहिजे etc. It should always precede with a past tense auxiliary such as पाहिजे होते (present, future) but never be preceded with an aspect marker or a present tense or future tense marker.

b) An infinitive must be followed by a mood or a tense marker when a modal auxiliary is absent, as in खाऊ देत असे or खायला लागत असे .

c) The modal auxiliary can be followed by a habitual aspect marker or by a subjunctive marker as depicted in the following examples.

- खाऊ शकतआहे (progressive)
- खाऊ शकला आहे (perfective)
- खाऊन बसले होते (completive)
- खाऊ शकतो (habitual)
- खाऊ शकले (subjunctive)

d) Infinitive marker, all aspectual markers, past tense auxiliary, conditional mood marker and future marker must be followed a gender-number marker.

XII. PROCEDURE FOR VG IDENTIFICATION

A VG is identified by scanning the sentence from left to right. The regular expression given below is used to detect VGs in a given sentence.

Star_Marker (Intermediate_marker) Must_end_marker*

There may be 0 or more occurrences of Intermediate_marker in a VG. The VG identifier uses the root, suffixes and the morphological features supplied by the morphological analyser and the POS tags assigned by the POS tagger. A VG begins as soon as a verb is scanned. The identified verb root may be locally POS ambiguous, i.e., noun or verb (eg. जेवण (noun) / जेवणे -'to eat' (verb)) since the root word 'जेवण ' is the same. The appropriate tag is selected by applying the regular expression on the verbal morphemes. If the sequence of the markers is allowed by the expression, they are included in the VG. The identification continues until a *must-end* marker is encountered. Once the end of the VG is marked, the group members are assigned fresh, disambiguated POS tags. The head of the VG is assigned VM while the auxiliaries are assigned VAUX along with the TAM features that they express.

XIII. TYPES OF POS AMBIGUITY IN VGS

a. Main Verb or Auxiliary Verb

In the sentence बघतच बसला होता (kept on seeing) बघतच appears as the progressive aspectual auxiliary as well as a main verb. Often a POS tagger is unable to resolve this ambiguity in the absence of contextual information.

b. Main Verb or Noun

In the sentence करून टाकले होते (had done), कर can appear as a noun (कर – tax, as in जकात कर वसूल केला) or as root of a verb कर(to do as in तो नियमीत व्यायाम करीत असे). In order to resolve this POS ambiguity, the system requires information that the auxiliary verb (असे) is preceded by the main verb. This information excludes the possible tag Noun and leaves Main Verb as the correct one. For example, in the sentence पावसामुळे क्रिकेटची मैच ३८-३८ ओव्हर ची करण्यात आली होती (Due to rain, the cricket match has been made of 48-48 overs), the verb group is identified as (करण्यात आली होती). (कर) is marked as a verb. Where as in the sentence जकात कर दिला गेला होता , it appears as a noun.

XIV. PERFORMANCE EVALUATION

In grammatically correct Marathi sentences , application of these hand-crafted rules to identify NPs and VPs proved to be almost 93.6% correct.

The total number of Marathi sentences on which the above method was manually tested was 3180. The minimum number of words in the sentence was 3 and the maximum 9. The number of unique (non-repeated words across sentences) words were 21060. The Marathi dictionary contains 51180 words.

We observed that long distance dependencies amongst words were difficult to handle using the above methods of identifying NGs and VGs. For example consider the following sentences.

- त्यांचे आसेपण

म्हणणे असतकि . . . . . .

- त्यांचे आसेपण म्हणणे पडतकि . . . . . .

The verb group is identified as (म्हणणे असत), whereas म्हणणे is a noun which should co-occur with the preceding possessive. But the possessive pronoun (त्यांचे) is not adjacent to (म्हणणे). Hence, the VG becomes incorrect.

XV. CONCLUSION

We have presented an algorithm to identify Marathi Noun groups by using the CD representation and morphological rules. NG and VG markers were discussed for chunking and the constraints that apply on the markers. A study of the grammatical categories and their markers that may appear inside a group were studied and hand-crafted rules were designed. Major POS ambiguities were resolved using these hand-crafted rules and the CD Parser. In Parsing or language generation these groups will be of importance. We cannot handle all the POS ambiguous cases (that involve scrambling or those that are structurally ambiguous) where immediate contextual rules do not help. However, using the ordering among the major categories and their possible combinations, we have tried to present ways that can be applied to other languages equally well. The methods are especially beneficial for languages with meagre corpora or other NLP resources. Since a system will not be able to learn patterns that might be absent in small training corpora, with the use of morphological patterns that govern the ordering of the elements inside a group, a large number of ambiguities and errors may be avoided at a first pass.

FUTURE WORK

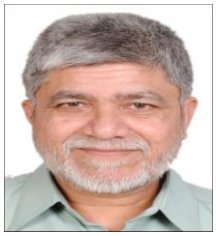
The work done so far in chunking Marathi text is done using a hand-crafted dependency parser for CD structure and using hand-crafted regular expressions and syntax rules. An automated system to represent the regular expressions and syntax rules and their application to the input Marathi text need to be considered. A dependency parser for CD needs to be designed so that it can be incorporated in the Chunking process.

REFERENCES

1. Abney, S., “Parsing by Chunks.” In Principle-Based Parsing, eds. B. Berwick, S. Abney, and C. Tenny, 257-278. Dordrecht: Kluwer Academic Publishers.
2. Bellaris H. S. K and Askhedkar L. Y, *A Grammar of the Marathi Language*, 1868, Obtained in digital format digitized by the Internet Archive in 2007 with funding from Microsoft Corporation.
3. Bharati, A., Chaitanya V., Sangal R, “Natural Language Processing: A Paninian Perspective.” New Delhi: Prentice-Hall of India, 1995.
4. Chakrabarti, D., Mandalia H., Priya R., Sarma V. and Bhattacharyya P., “Hindi Compound Verbs and their Automatic Extraction”, In Proc. of Computational Linguistics Conference (COLING), Manchester, UK, 2008.
5. Chakrabarty, D., Sarma V., Bhattacharyya P., "Complex Predicates in Indian Language Wordnets", Lexical Resources and Evaluation Journal, 40 (3-4), 2007.
6. Dalal, A., Nagaraj K., Sawant U., Shelke S., “Hindi Part-of-Speech Tagging and Chunking: A Maximum Entropy Approach.” In the Proceedings of the NLPAI Machine Learning Workshop on Part Of Speech and Chunking for Indian Languages. Mumbai, India, 2006.
7. Deshpande Madhuri, Gore Sharad, "Using Conceptual Dependency Theory to represent Marathi text", International Journal of Computer Technology & Applications, Vol 6 (6), pp. 911-914, 2015.
8. Elaine Rich, Kevin Knight, Shivashankar Nair. *Artificial Intelligence*, 2009, Third Edition, Tata-McGraw Hill Education Private Ltd, ISBN 978-0-07-008770-5, ch. 10.
9. Gune H., Bapat M., Khapra M., Bhattacharyya P., “Verbs are where all the Action Lies: Experiences of Shallow Parsing of a Morphologically Rich Language”, Computational Linguistics Conference (COLING), Beijing, China, 2010.
10. Singh, A., Bendre S. M., Sangal R., “HMM Based Chunker for Hindi.” In the Proceedings of International Joint Conference on NLP, 2005.
11. Singh, S., Gupta K., Shrivastava M. and Bhattacharyya P. (2006). “Morphological Richness Offsets Resource Demand – Experiences in Constructing a POS Tagger for Hindi.” In the Proceedings of the COLING/ACL-2006, 779-786. Sydney, Australia, July.
12. Smriti Singh, Om P. Damani, Vaijayanthi M. Sarma, "Noun Group and Verb Group Identification for Hindi", COLING, December, 2012.
13. Steven Lytinen, "Conceptual Dependency and its Descendents.", Computer Math. Applic, 23(2-5): 51-73, 1992.
14. R. Schank, "Conceptual dependency: A theory of natural language understanding", Cognitive Psychology 3, 552-631, 1972.
15. R. Schank and R. Abelson, *Scripts, Plans, Goals, and Understanding*, Lawrence Erlbaum Associates, Hilladale, NJ, 1977.
16. Ray, P. R., Harish V., Basu A., Sarkar S., “Part of Speech Tagging and Local Word Grouping Techniques for Natural Language Parsing in Hindi.” In the Proceedings of (ICON). Mysore, India, 2003.
17. Ramshaw, L. A., Mitchell, P. M., “Text Chunking Using Transformation-Based Learning.” In the Proceedings of the Third ACL Workshop on Very Large Corpora, 82-94. Cambridge, MA, USA, 1995.
18. Roger C. Shank, Larry Tesler, "A Conceptual Dependency Parser For Natural Language", International Conference on Computational Linguistics (COLING), Sweden, September 1969.
19. Walimbe M. R., *Sugam Marathi Vykarana Lekhan*, Nitin Prakashana, Pune. ISBN 81-86169-80-6, 2012.
20. Wlodek Zadrozny, Karen Jensew, "Semantics of Paragraphs", Computational Linguistics, Volume 17, Number 2 Pg 171-209, 1991.
21. http://www.tdil-dc.in

AUTHOR PROFILE

Mrs. Deshpande Madhuri M. , B.C.S, M.C.S, M.Phil (CS), a research student in the Department of Computer Science, Savitribai Phule Pune University (formerly University of Pune), also a member of Computer society of India (CSI). Research interests include areas of Marathi Language processing using data mining tools, statistical tools and artificial intelligence.



Dr. Gore Sharad D., has worked in the Department of Statistics, University of Pune (now Savitribai Phule Pune University) from 1978 to 2014. He headed the Department of Computer Science (2001-2004) and Statistics Department (2009-2012). He also taught at Pennsylvania State University (1990-1993) and Bharati Vidyapeeth Deemed University's Institute of Management and Entrepreneurship Development (IMED) (2006-2007). He has written 3 books, published 70 research papers in International Journals, and participated in more than 100 National and International Conferences. He has guided Ph. D. students in Statistics, Computer Science, and Environmental Sciences.