Chapter 6

Approaches for Automatic Link Acquisition

6.1 Introduction

Parsing link grammar of any language may it be English or Marathi needs some dictionaries where information such as root word, various forms of words, links associated with it etc to be kept in databases. Meaning, link grammar framework need language specific information for parsing. At present acquiring this data is done manually with the help of linguists or lexicographers. Some semiautomatic approach is also introduced in recent developments. But to acquire such data automatically is still a challenging task. Recent developments in linguistic resources makes availability of WordNet [Miller, 1995], VerbNet [Kipper-Schuler, 2005] etc. Using such resources in specific applications in current system for improving effectiveness is also one of the challenge.

Link grammar formalism establishes links between any two words by looking at their associations or agreement rules with each other. In particular, Marathi link grammar parsing system searches for links between verbs to other nominal present in the sentence. Ideally system should return correct Karaka relation or any other relation appropriate between the pair of words. We have observed that in few cases nominal is not present even in databases, treating them as an unknown word. This hampers
whole process of parsing. In such case if we introduce a system which acquires link among two words automatically be of great use in improving parsing effectiveness. We propose two approaches to acquire links between verb and other nominal in a sentence by using Marathi WordNet and Named Entity Recognition.

6.2 WordNet Approach

WordNet is a large semantic network interlinking words and groups of words by means of lexical and conceptual relations represented by labeled arcs. WordNet’s building blocks are synonym sets, unordered sets of cognitively synonymous words and phrases. Each member of a given synset expresses the same concept, though not all synset members are interchangeable in all contexts. Examples are Car, Automobile, hit, strike and big, large. All synsets further contain of four separate components, each containing synsets with words from the major, open class, syntactic categories nouns, verbs, adjectives and adverbs. Morphosemantic relations link words from all four parts of speech that are both morphologically and semantically related.

Princeton WordNet [Miller, 1995] is a lexical knowledge network of English words. Its growing popularity as a useful resource for English and its incorporation in natural language tasks has prompted the creation of similar WordNets in other languages as well.

Hindi WordNet is inspired by this famous English WordNet. Hindi and Marathi languages belong to the same language family. The close correspondence between Hindi and Marathi reduced to programmatically borrowing lexico-semantic relations from the Hindi WordNet into Marathi WordNet since the Marathi WordNet is largely aligned with the Hindi WordNet. Since 2002 experiments and development in indo WordNet are taking place as a result Marathi WordNet 3.0\(^1\) is available for downloading under GNU-GPL.

We need to understand the entries in Hindi or Marathi or any Indian WordNet which

\(^1\)http://www.cfilt.iitb.ac.in/wordnet/webmwn/index.php
consists following elements,

1. Synset - Set of Synonymous words according to frequency of usage of a particular word, they are arranged in the synset.

2. Gloss - It describes the concept by giving two parts: Text definition and Example sentence, which gives the usage of the words in the sentence.

3. Position in Ontology - An ontology is a hierarchical organization of concepts, more specifically, a categorization of entities and actions. For each syntactic category namely noun, verb, adjective and adverb a separate ontological hierarchy is present. Each synset is mapped into some place in the ontology. A synset may have multiple parents.

A WordNet is a word sense network. A word sense node in this network is a synset which is regarded as a basic object in a WordNet. Each synset in the Hindi WordNet is linked with other synsets through the well known lexical and semantic relations of hypernymy, hyponymy, meronymy, troponymy, autonymy, entailment etc. Semantic relations are between synsets and lexical relations are between words. WordNet also gives relations between the synsets of different parts of speech i.e. linkage between nominal and verbal concepts, linkage between nominal and adjectival concepts and linkage between adverbial and verbal concepts.

6.2.1 Proposed Approach

WordNet gives semantic as well as lexical relations and is able to give relation between the synsets of different parts of speech too. Our main objective to propose this approach is to identify links between two words for a input sentence automatically by using Marathi WordNet. Relation links available through Marathi WordNet will be useful in achieving this objective. Following approach is suggested,

1. Take a training data set around 50-100 sentences. Parse them and restore linking scheme gained from parsing.
2. Draw ontological nodes of each words of a sentence from Marathi Wordnet.

3. Sentence by Sentence analyze ontological nodes found in pair from verb to other POS found in WordNet i.e. Noun, Adjective, Adverb.

4. Identify parameters by which the ontological nodes of these pair matches with each other.

5. Formulate heuristic to find link between any unknown pair of verb to Noun, Adjective, Adverb by using it.

The proposed approach right now is in rough form but it is certainly going to provide some direction in identifying links automatically using Marathi WordNet.

To brief this idea let us have a look in following example, Ontological node of word *vyakti* is shown in Figure 6.1,

below, *bolane* is the action word associated with *vyakti*. For *bolane* we get entry as shown in Figure 6.2 in Marathi WordNet,

Among all the sense tree system needs to pick the appropriate candidate tree by checking other words in input sentence. From verb to other nominal words association...
Figure 6.2: Word Sense List of bolane in Marathi WordNet
score has to be calculated. It is either by statistical method or by a heuristic function based on algorithms like Apriori etc.

If suppose following sense is selected then ontology of it will be explored as following Figure 6.3.

Figure 6.3: Ontological Node of *bolane* Explored in Marathi WordNet

The approach proposed that is acquisition of links automatically for any language’s link grammar formulation purpose is novel. It needs lots of efforts to put this idea into workable framework. One may consider this as an future extension to the current work.

### 6.3 Named Entity Recognition Approach

Named Entity Recognition (NER hereafter) is the task of identifying and classifying people, organization and other named entities within text. NER is central to many NLP systems, especially Information extraction tasks, people noticed that it is essential to recognize information units such as names, including person, organization and location names and numeric expression including time, date, money and percentage. Identifying references to these entities it text was acknowledged as one of Information extraction’s important subtasks. Before NER field was recognized in 1996, significantly research was conducted by extracting proper names from texts. A paper published in 1991 by Lisa Rau is often cited as the root of the field [Rau and Jacobs, 1991]. The vast majority of proposed system fall in two categories: the handmade rule based systems
and the supervised learning based systems. In both approaches large collections of
documents are analyzed by hand to obtain sufficient knowledge for designing rules or
for feeding machine learning algorithms. Expert linguists must execute this important
amount of work, which in turn limits the building and maintaining large scale NER
systems.

Both the methods mentioned above require the efforts of language expert. An approxi-
mately large set of annotated data is yet to be made available for the Indian languages.
There has been very little work in the area of NER in Indian languages. Although much
work has been accomplished for English language in NER using techniques such SVF,
CRF and HMM, adaption of such techniques in Indian languages is also noticeable.
The statistical conditional random field model has been used to develop NER system
of many Indian languages like Hindi, Bengali, Oriya, Urdu, Telugu etc. [Ekbal et al.,
2008].

The training data has been learned for many types of Named Entities, Following is the
list of NE identified through various literature stated above,

- Named Entity Person
- Named Entity Location
- Named Entity Organization
- Named Entity Time Expression
- Named Entity Measurement
- Named Entity Artifacts
- Named Entity cuisines
- Named Entity Entertainment
- Named Entity Disease
- Named Entity Plants etc.
Some more Named Entities were also handled by systems that have specific needs like ‘film’, ‘Scientist’, ‘Book title’, ‘Project Title’ etc...

6.3.1 Proposed Approach

The objective of proposing this approach is to optimize the Link dictionary used for Link Grammar Parsing by using NER systems. It is quite common that a parsing texts may involve such named entities discussed in above section. If the parsing system would be able to get identified before look up phase for such named entities to parse, it will definitely reduce time of parsing and will result in optimum entries in the databases too. Following steps can be taken to acquire links automatically using NER,

1. Tag the input sentence with appropriate Named Entity if exists.

2. For the list of Named Entity present check whether a link(s) for verb to Named Entity in dictionary, exists or not.

3. Build links if found any such link according to it.

The steps discussed above have subtasks to be performed. The first step of course is a task of NER in a input sentence given for parsing. One can use effective NER system as a preprocessing step. Step two involves screening of word and link dictionary. Following example will give brief idea of it.

The input sentence is ‘lili ne kaanda kapla’ suppose our link dictionary have link kapane (verb) to bhaji kand (Noun) as.

Our proposed system will follow the steps discussed above,

Step 1 : Tag input sentence for Named Entity

lili /NEP ne kaanda /NEV Kapla

NEP - Named Entity Person
NEV - Named Entity Vegetable
**Step 2:** In this step, verb to nominal links will be built as per Marathi LG Parsing.

Resulting as follows,

\[ \text{Kapane} \rightarrow \text{NEP with pratyaya ne} \]

\[ \text{Kapane} \rightarrow \text{NEV with pratyaya } \Phi \]

NER system should have predefined Named Entity Vegetables list consisting all Marathi names used for vegetables i.e.

- methi
- tamata
- kaanda
- mula
- gajar
- bhedi
- vangi etc

If we would be able to make our system to store links in this format we will achieve optimization in it. Whenever next sentence encounters mini ne gajar kapale system need not have separate link for it. This proposed approach depends upon linguistic resource like NE tag sets, NE lists and effective algorithm to tag input sentence for NE, While parsing the look up phase and associating links with appropriate named entity. Certainly it will be of great help to achieve effective results of Marathi Link Grammar Parsing.

### 6.4 Concluding Remarks

In this chapter we have proposed two approaches of automatic link acquisition by using Marathi WordNet and Named Entity Recognition Systems. Parsing in a link grammar
framework identifies links between words of a input sentence. If we would be able to build such links based on such available resources, then it would be of great help and will be able to obtain effective parsing results.