Morphological Analysis for a given text In Marathi language

[1]Aditi Muley,[2]Manaswi pajai, [3]PriyankaManwar ,[4]Sonal Pohankar,[5]Gauri Dhopavkar

Department of Computer Technology, YCCE

Nagpur- 441110, Maharashtra, India

[1].aaditi.muley@gmail.com,[2]manaswipajai11@gmail.com

[3]priyankasmanwar@gmail.com,[4]sonalpohankar1993@gmail.com

[5]gauri.manoj@gmail.com

Abstract

Morphology is the field of the linguistics that studies the internal structure of the words. Morphological Analysis and generation are essential steps in any NLP Application. Morphological analysis means taking a word as input and identifying their stems and affixes. Morphological Analysis provides information about a word's semantics and the syntactic role it plays in a sentence. Morphological Analysis is essential for Marathi as it has a rich system of inflectional morphology as like other Indo-Aryan family languages. Morphological Analyzer for analyzing the given word and generator for generating word given the stem and its features (like affixes). This paper presents the morphological analysis for Marathi Language using Ruled Bases Approach. This project has been developed to find a root word of a given word and can be used in Gender Recognition as well.

1) INTRODUCTION

1.1. NLP (Natural language Processing)

Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. In this paper, we present the morphologicalanalyzer for Marathi which is official language of the state of Maharashtra (India). With 90 million fluent speakers worldwide, Marathi ranks as the 4th most spoken language in Indiaand the 15th most in the world. [1]

1.2 Marathi morphology

In linguistics, morphology is the identification, analysis and description of the structure of a given language's morphemes and other linguistic units, such as words, affixes, parts of speech, intonation/stress, or implied context. Morphological typology represents a method for classifying languages according to the ways by which morphemes are used in a language from the analytic that use only isolated morphemes, through the agglutinative ("stuck-together") and fusional languages that use bound morphemes (affixes), up to the polysynthetic, which compress lots of separate morphemes into single words. While words are generally accepted as being (with clitics) the smallest units of syntax, it is clear that in most languages, if not all, words can be related to other words by rules (grammars). For example, English

speakers recognize that the words dog and dogs are closely related differentiated only by the plurality morpheme "-s", which is only found bound to nouns, and is never separate. Speakers of English (a fusional language) recognize these relations from their tacit knowledge of the rules of word formation in English. [2]

1.2 The Alphabets

Marathi script consists of 16 vowels and 36 consonents making a total of 52 alphabets.

1.3 Vowels

The vowels are grouped in two groups. The first group consists of 12 vowels as follows: aaa(A) i ii(I) u uu(U) e ai o au aMaH The first 10 vowels are very widely used. The last two are less commonly used. Suffix stripping is a pre-processing step required in a number of natural language processing applications such as information retrieval, text summarization, document clustering, and word sense disambiguation.The stem is not necessarily the linguistic root of the word. Earlier work in this direction for Indian languages includes Hindi, Bengali, Tamil, and Oriya. But very little amount of work has been done for Western Indian languages like Marathi and Konkani.[2]

2) Motivation and Problem Definition

A highly inflectional language has the capability of generating hundreds of words from a single root. Hence, morphological analysis is vital for high level applications to understand various words in the language. Morphological analyzer forms the foundation for applications like information retrieval, POS tagging, chunking and ultimately the machine translation. Morphological analyzers for various languages have been studied and developed for years.Eryiğit and Adalı (2004) propose a suffix stripping approach for Turkish. The rule based and agglutinative nature of Turkish allows the language to be modeled using FSMs and does not need a lexicon. The morphological analyzer does not face the problem of the changes taking place at morpheme boundaries which is not the case with inflectional languages. Hence, although apprehensible this model is not sufficient for handling the morphology of Marathi. Our problem definition is root word and gender analysis for a given text in Marathi language. In this paper we are going to see how the root word of a given word is found and recognises the Gender of the sentence.[3]

3) LITERATURE SURVEY

In 2001, Shambhavietal introduced Kannada morphology analyzer and generator and using tire [11]. A lightweight stemmer for Hindi [12] was developed by Ramanathan et al. in the year of 2004. In this research, words conflate terms by suffix removal for information retrieval. Willet.P proposed the porter stemming algorithm for electronic library and information system [13] in 2006. Zahurul.MD et al. developed a lightweight stemmer for Bengali [14] in the year of 2009 for Bengali language spell checker. Assas-band, an affix exception list based Urdu stemmer [15] was developed by Qurat-Ul-Ain Akram and etal. in the year of 2009. It stems the Urdu words using lexical lookup method (Assas-band). In 2010, Dinesh Kumar and Prince Rana developed design and development of stemmer for Punjabi [16], it uses Brute Force algorithm for stemming the Punjabi words.VijaySundar et al. introduced Malayalam stemmer for information retrieval [17] in the year of 2010. Finite State Automata method is used to stem the Malayalam words.

4) ARCHITECTURE AND DESIGN

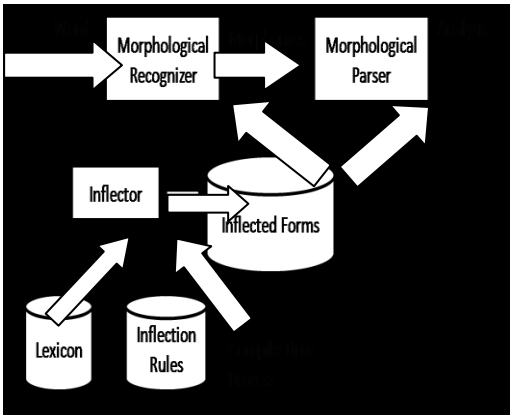

Figure 4.1 Architecture of Marathi Morphological Analyzer

4.1 Morphological Analyzer for Marathi

The formation of polymorphemic words leads to complexities which need to be handled during the analysis process.

4.2 Linguistic Resources

The linguistic resources required by the morphological analyzer include a lexicon and inflection rules for all paradigms.[4]

4.2.1 Lexicon

An entry in lexicon consists of a tuple <root,paradigm, category>. The *category* specifies the grammatical category of the root and the paradigm helps in retrieving the inflection rules associated with it. Our lexicon contains in all 24035 roots belonging to different categories.

4.2.2 Inflection Rules

Inflection rules specify the inflectional suffixes to be inserted (or deleted) to (or from) different positions in the root to get its inflected form. An inflectional rule has the format: <inflectionalsuffixes ,morphosyntactic features, label>. The element morphosyntactic features specifies the set of morphosyntactic features associated with the inflectional form obtained by applying the given inflection rule. Following is the exhaustive list of morphosyntactic features to which different morphemes get inflected:

1) Gender: Masculine, Feminine, Neuter, Common.

2) Number: Singular, Plural, Non-specific

3) Tense: Past, Present,Future

5) Implementation Methodology

Algorithm for Root word Analysis:

Input: List of words

Output: Stem of words

Step 1: Eliminate all the complex suffixes.

e.g:- कडून , मुळे, साठी , प्रमाणे, वरून, वर ,...

Step 2: Eliminate the join word suffixes i.e. Eliminate the inflections of consonants like च, ल, ण, ळ, र, द... with या

e.g., द्या =<द> + <या>

Step 3A: Eliminate the inflections for consonant च

Step 3B: Eliminate the inflections for consonant ल

Step 3C : Eliminate the inflections involving plain suffixes

Gender recognition for a given text in Marathi language:

As in Gender recognition we use the format of (SOV) Subject ,Object and Verb,we first check for subject.If the subject matches with the database then we get the result.If subject is same for both genders then it checks for verb and thus the result is obtained.

Following are some of the examples of Gender recognition:

1) तोघरीज़ातो.

In this example we recognize the gender first by subject. As subject matches with the database we get the result as Masculine Gender.

2)मीशाळेतजाते.

In this example as we are unable to recognise the Gender by the subject so we recognise the gender by verb. As verb matches with the database we get the result as Feminine Gender.

3)आम्हीबाहेरज़ातो.

In this example as we are unable to recognise the Gender by the subject we recognise the gender by verb. As verb matches with the database we get the result as Neuter Gender.

6) Experimental Results

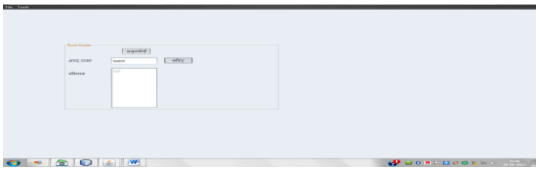

Figure 6.2 Output for Root Word Analysis

In this snapshot we have entered the input for which the root word is to be found. We have given the input as घरात so we get the output as घर which is the root word for the input .Similarly other examples for which the Root word can be found are as follows:

1. देशावरदे⟹

   In this example the root word is देश.

2.घरातघर⟹

In this example the root word is घर.

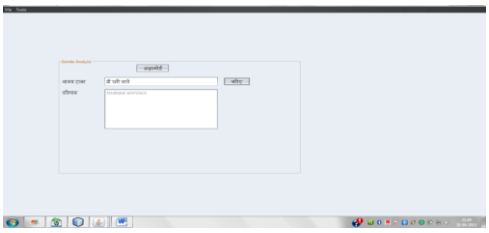

Figure 6.4 Result of Feminine Gender

In this snapshot we have entered the input for which the Gender is to be recognised. We have given the input as तीघरीज़ाते. In this example we first check the subject. As the subject matches with the database of Feminine Gender we get the output as Feminine Gender.

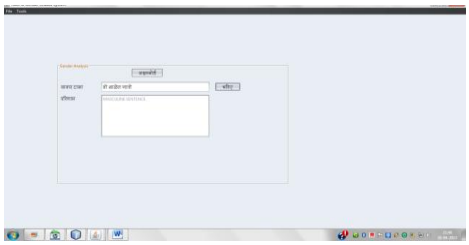

Figure 6.4 Masculine Gender

In this snapshot we have entered the input for which the Gender is to be recognised. We have given the input as तोशाळेतज़ातो. This example we first check the subject. As the subject matches with the database of Masculine Gender we get the output as Masculine Gender.

7) CONCLUSION

Thus we conclude that morphological processing improves the retrieval performance for Marathi Language. Thus more attention has to be given to morphological analyzer. Also effect of stop-words on information retrieval is observed. An important observation is that the suffixes in Marathi can also contribute to the semantics of the document and hence improves the retrieval performance. The current morphological analyser does not handle derivational morphology. In Marathi, derivational morphology is a very productive way of forming words. Handling derivational morphology can also increase the system performance. Foreign words (transliterated English words in Marathi text) can be stemmed heuristically to improve the performance of the system. We presented a high accuracy morphological analyzer for Marathi which very efficiently finds the Root word of a given word and

recognises the Gender of the sentence which the use inputs.

8) REFERENCES

[1]GaganBansal, Satinder Pal Ahuja, Sanjeev Kumar Sharma,"Improving Existing Punjabi Morphological Analyzer",Research Cell: An International Journal of Engineering Sciences ISSN: 2229-6913 Issue Dec. 2011, Vol. 5.

[2]MugdhaBapat , HarshadaGune,Pushpak Bhattacharyya,"A Paradigm-Based Finite State Morphological Analyzer for Marathi", Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing (WSSANLP), pages 26–34, the 23rd International Conference on Computational Linguistics (COLING), Beijing, August 2010.

[3]Oflazer,Kemal."Two-level Description of Turkish Morphology". InTheEuropeanChapter of the ACL (EACL).

[4]Raj Dabre, ArchanaAmberkar, Pushpak Bhattacharyya. 2012. "Morphological Analyzer for Affix Stacking Languages: A Case Study of Marathi", Conference on Computational Linguistics (COL-ING).

[5] Kannada Morphological Analyzer and Generator Using Triepaper.ijcsns.org/07_book

[6]A.Ramanathan and D.Rao, "A Lightweight Stemmer for Hindi ," in
proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics(EACL) on Computational
linguistics for South Asian Language (Budapest, April) workshop,
2003.

[7] The Porter Stemming Algorithm: Then and Now - White Rose, eprints.whiterose.ac.uk**,** 1434/01 willettp9_PorterStemmingReview.pdf

[8] Khan. 2007. "A light weight stemmer for Bengali and its Use in spelling Checker," Proc. 1st Intl. Conf. on Digital Comm. and Computer
Applications (DCCA07), Irbid, Jordan, March 19-23.

[9] Assas-Band, an affix-exception-list basedUrdustemmer**,**dl.acm.org/citation.cfm.

[10] Hybrid Approach for Stemming in Punjabi - International Journal of
Computer Science and Computer Network, www.ijcscn.com,
ijcscn2013030206.pdf

[11]Malayalam Stemmer - Computational Linguistic Research Group, nlp.aukbc.org, Malayalam Stemmer.

[12]M.Thangarasu, Dr.R.Manavalan,"A Literature Review: Stemming Algorithms for Indian Languages".International Journal of Computer Trends and Technology (IJCTT) – volume 4 Issue 8–August 2013