Marathi Derivational Morphological Analyzer

Yash Waghmare[1], Aditi Kale[2], Gauri M. Dhopavkar[3]

[1,2]Students Yeshwantrao chavan College of Engineering, Nagpur,India,441110
[3] Professor Yeshwantrao chavan College of Engineering, Nagpur,India, 441110

Abstract – *Marathi is an Indian language which is relatively rich in morphology. Words can be formed from a number of morphological operations, the commonest being inflection, derivation and compound. Morphological analyzers should be able to handle these processes, especially if they occur frequently in the language. While most morphological analysers can tackle inflectional operations easily, derivation is harder because of its less regular nature. In this paper,we present our Marathi derivational morphological analyzer. This approach increases the coverage of existing morphological dictionary.*

Keywords- morphology, derivational, inflectional Morphological analyzers, morphological dictionary.

1 INTRODUCTION

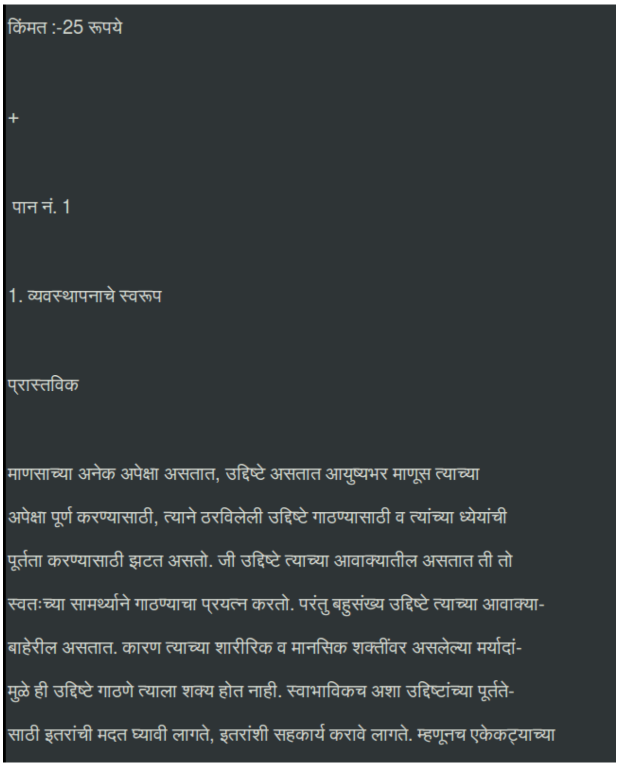
Morphology is the study of processes of word formation and also the linguistic units such as morphemes, affixes in a given language. It consists of two branches: derivational morphology and inflectional morphology. Derivational morphology is the study of those processes of word formation where new words are formed from the existing stems through the addition of morphemes. The meaning of the resultant new word is different from the original word and it often belongs to a different syntactic category. Example: happiness (noun) = happy (adjective) + ness. Inflectional morphology is the study of those processes of word formation where various inflectional forms are formed from the existing stems. Number is an example of inflectional morphology. Example: cars = car + plural affix 's'. The main objective of our work is to develop a tool which executes the derivational morphological analysis of Marathi. Morphological analysis is an important step for any linguistically informed natural language processing task. Most morphological analyzers perform only inflectional analysis. However, derivational analysis is also crucial for better performance of several systems. Moreover, an output that is richer in morphological details is useful, especially in Machine Translation between two closely-related languages. They are also used in search engines to improve the information extraction. [1] Since derivational processes can often be productive in a language, the development of an effective derivational analyzer will prove beneficial in several aspects.

We developed a Derivational Morphological analyzer for Marathi over a XML Paradigm based analyzer.

2 PROPOSED METHODOLOGY

2.1 Data Cleaning

Data cleaning is the process of detecting and correcting or removing, corrupt or inaccurate data. It refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. In computing, plain text is a loose term for data that represent only characters of readable material . It may also include a limited number of characters that control simple arrangement of text, such as spaces, line breaks, or tabulation characters. It is important to remove those punctuations, special characters, number and stopwords should be removed. Figure 1 is an example of type of input data.



किंमत :-25 रूपये

+

पान नं. 1

1. व्यवस्थापनाचे स्वरूप

प्रास्तविक

माणसाच्या अनेक अपेक्षा असतात, उद्दिष्टे असतात आयुष्यभर माणूस त्याच्या
अपेक्षा पूर्ण करण्यासाठी, त्याने ठरविलेली उद्दिष्टे गाठण्यासाठी व त्यांच्या ध्येयांची
पूर्तता करण्यासाठी झटत असतो. जी उद्दिष्टे त्याच्या आवाक्यातील असतात ती तो
स्वतःच्या सामर्थ्याने गाठण्याचा प्रयत्न करतो. परंतु बहुसंख्य उद्दिष्टे त्याच्या आवाक्या-
बाहेरील असतात. कारण त्याच्या शारीरिक व मानसिक शक्तींवर असलेल्या मर्यादां-
मुळे ही उद्दिष्टे गाठणे त्याला शक्य होत नाही. स्वाभाविकच अशा उद्दिष्टांच्या पूर्तते-
साठी इतरांची मदत घ्यावी लागते, इतरांशी सहकार्य करावे लागते. म्हणूनच एकेकट्याच्या

Fig. 1- Sample Input Data

2.2 Using Lttoolbox

Lttoolbox is an open-source finite-state toolkit used for morphological analysis, requires the creation of a morphological dictionary that shows correspondences between surface forms and lexical forms[2]. Surface Forms are the inflected forms of words that would be found in texts whereas Lexical Forms refer to the base forms of those words. For instance, the word gAvAlA (village +DAT) is the Surface Form of the Lexical Form gAva + lA. This mapping allows the finite state transducer to process the stream of morphemes correctly, depending on whether it is analysis or generation. Lttoolbox allows the user to do both. The analyser will take a Surface Form as input to return the Lexical Form and vice versa for the generator. The regularities seen in the correspondences between Surface Form and Lexical Form are easily encoded in the form of paradigms. The paradigms are actually rules that are organized in blocks. [3] A group of words that belong to one paradigm will follow the same set of spelling rules and take the same kind of affixes. A paradigm is created in the morphological dictionary file using the XML format described in the toolkit. Hence, within a paradigm, entry <e> encloses the correspondence between the elements <l> and <r> which stand for the left element and the right element respectively. The tag <p> includes both these left and right elements. The correspondence shows the transformation that will take place when analysis takes place. Hence, in Fig 1 Surface Form raswyAlA will result in the analysis of raswA as the root form with the features enclosed in the <r> element.

```
<pardef n =“rasw/A n”>
<e>
<p>
 <l>A</l>
<r>A<sn = n/><sn = sg/><sn = parsarg:0/></r>
 </p>
 </e>
 <e>
 <p>
 <l>yAlA</l>
<r>A<sn = n/><sn = sg/><sn= parsarg:lA/></r>
 </p>
 </e>
</pardef>
```

Fig. 2- Inflectional Paradigm for Marathi raswA in the Lttoolbox format.

2.3 Paradigms for Derived Forms

If we reconsider some of the characteristics of derivation, the following points are worth noting:

- Change in the category and meaning of a word after derivation
- Less regular in nature than inflection
- Operates along with inflection

The first of these characteristics, change in the category and meaning of a word is done by specifying the grammatical features in the morphological dictionary. The second problem, that of its less regular nature is solved by listing out the root words under paradigms created specifically for derivational affixes. As this is a database driven method, listing the roots is the only solution. However, once this resource has been created, further experiments can be carried out to find criteria for attachment that are easier to model. Finally, we have the morphological operations of inflection and derivation that operate one after the other. This is a more difficult problem to tackle while analyzing a word as it must be segmented according to more than one affix. While tackling this aspect of derivational morphology, the Split

Morphology Hypothesis [3] is taken into consideration. This says that inflection is usually the last operation to take place, i.e., it is never followed by derivation. According to this assumption, words could take affixes in the following order.

1. stem + inflectional suffix

2. stem + derivational suffix

3. stem + derivational suffix + inflectional suffix

4. stem + derivational suffix + derivational suffix + inflectional suffix

The sections that follow will expand upon the paradigm forms that have been created for Marathi in the morphological dictionary. The derivational paradigms are a layer built in addition to the existing inflectional paradigms in order to deal with both operations at once. These Paradigms can be used to handle cases of derivation, cliticization etc. We can also create a nested Paradigm, meta Paradigm to handle above mentioned cases.[4]

2.4 Studying Marathi Derivations

To build the derivational morphological analyzer,we conducted a study to identify the derivational suffixes and the related morphological changes. After identifying the suffixes, the rules pertaining to these suffixes were obtained.[5]

All the word entries in the dictionary were studied. The study of words helped us in identifying some of the derivational suffixes present in the dictionary. For example let us consider the word AyuSaBara. This word is derived from the word AyuSa (AyuSaBara = AyuSa (life)+Bara). But Bara cannot be confirmed as a suffix because of just one instance. In order to confirm Bara as a suffix, even other words ending with Bara must be examined. The more the number of words we find, the greater is the productivity of the suffix. Words like divasaBara(divasaBara = divasa + Bara) and SaNaBara (SaNaBara = SaNa + Bara). Only relevant words were studied and the suffixes were obtained only from them as there might be some exception. A set of rules will be developed for such derivational suffix. Table 2 shows a few suffixes and their derivations.

A set of suffix replacement rules and a dictionary in our analyzer having taken insights from Porter's stemmer [6] and K-stemmer[7]. Primary goal of Porter's stemmer is suffix stripping. It achieves task in 5 steps applying rules at each step. These concepts are used to identify stem.

Table 1- Example derivations of some suffixes

Sr. No.	Suffix	Root Word	Words
1	wun	KidkI	KidkIwun
2	xAra	xukAna	xukAnaxAra
3	cA	Aj	AjcA
4	lA	jANe	jAylA
5	wo	basNe	baswo

2.5 Finding Majority Properties

The majority properties (of derived words of a suffix) are the properties which most of the words exhibit. Example: let us consider the derived words of the suffix Bara. There are 36 derived words of the Bara suffix in the root word dictionary. Most of them follow AyuSa Paradigm Class. The majority properties of a suffix help us in the derivational analysis of the unknown derived words of that suffix.Thus the genuine derived words which are unknown to the analyzer will be analyzed using the majority properties. The majority properties of derived words were obtained in two main steps. First, a suffix was considered. Then all the derived words pertaining to that suffix were acquired. Genuine derivations were found out using the suffix derivational rules.

A suffix table was built using the majority properties of the derived words. The suffix table contains all the suffixes, respective Paradigm Class and their inflectional forms. Table 4 contains few suffixes and their inflectional forms with the name of their Paradigm Class. Example : let us take the word AyuRyaBara (ending with Bara). First, the normal form is obtained from suffix table and paradigm class. The word is present in the dictionary and it also satisfies genuineness property. Hence AyuRyaBara is accepted as a derived word. If it is not found in dictionary then it is removed.

Table 2- suffix table

Sr. No.	Suffix	Suffix- forms	Paradigm Class
1	wun	wun	Gara
2	xAra	xAra,xarI	karja
3	cA	cA,cI,ce	Sabda
4	lA	lA	gAva
5	wo	wo,we	Sika

The possible inflections of a suffix can be derived from its majority properties. This information was stored in a table. The majority properties and the suffix table play an important role in the analysis of the unknown words.

2.6 Using Wordnet Data for Confirming Genuineness of word

If an invalid word is not analyzed by the inflectional analyzer, there is no need for proceeding to the derivational analysis of that word. Therefore the genuineness of a word must be tested before going for the derivational analysis. The Wordnet is chosen as a resource that enables us to test the genuineness of a word. A word will be treated as a genuine word only when it is present in that list.

A WordNet is a lexical database for the Marathi language.

It groups English words into sets of synonyms called *synsets*, provides short definitions and usage examples, and records a number of relations among these synonym sets or their members. WordNet can thus be seen as a combination of dictionary and thesaurus. A list was generated from this data having 40k words approximately.

3 DESIGN

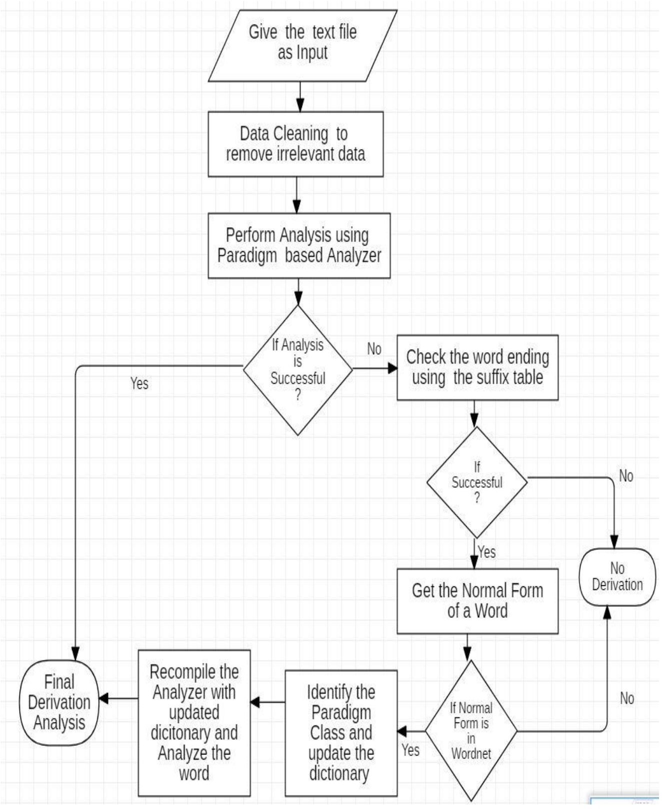

Give the text file as Input
Data Cleaning to remove irrelevant data
Perform Analysis using Paradigm based Analyzer
If Analysis is Successful ?
No
Yes
Check the word ending using the suffix table
If Successful ?
No
Yes
No Derivation
Get the Normal Form of a Word
If Normal Form is in Wordnet
No
Yes
Identify the Paradigm Class and update the dictionary
Recompile the Analyzer with updated dicitonary and Analyze the word
Final Derivation Analysis

Fig. 3 Algorithm

The input to the algorithm is a text. The output is a derivational analysis of the each word. For example: if the word is raswyAlA. First, the analyzer tries to analyze the input word. In this case the word raswyAlA is a noun, singular in number, etc. Then it gives the information (category, gender) of the root word (raswA) from which the input word is derived (derivational analysis). If the word is not present in the dictionary of analyzer then it’s added to the dictionary.

Examples

The following are 3 examples explain the working of an approach. These examples are provided to give a clear picture of the complete approach.

a)Example 1

Word : AyuRyaBara

First the word is analyzed using our morphological analyzer, if the word entry is present in Morphological

Dictionary then it will analyze the word “AyuSaBara” and print the output.

b)Example 2

Word : divasBara

The Analyzer cannot analyze this word. The word divasBara is ending with one of the forms (Bara) present in the suffix table. The normal form of the input word is obtained by replacing the suffix form in the input word with the suffix. Hence the normal-form of the input word divasBara is divasa. The word divasa is present in Wordnet data and paradigm class is obtained. Then it’s entry is added to morphological dictionary and then it gives it’s output properly.

c)Example 3

Word : KarcaBara (invalid word)

The analyzer cannot analyze this word. The word KarcaBara is ending with one of the forms (Bara) present in the suffix table. But the normal form of KarcaBara is not present in wordnet. So there is no derivational analysis for this particular case.

4 CONCLUSION

We presented an approach which updates existing morphological analyzer for performing derivational analysis. The approach uses the main principles of both the Porter’s stemmer and Krovetz stemmer for achieving the task.. It also expands the coverage of the derivational analyzer. But it must be incorporated in applications like machine translators which use derivational analysis for understanding its real strengths and limitations.

REFERENCES

1. *https://pdfs.semanticscholar.org/ed15/cbfeea18b4c592762461875d041f5f66cf59.pdf*
2. *M. L. Forcada, B. Bonev, Ortiz S.Rojas, and F. Tyers. 2007. Documentation of the open-source shallow-transfer machine translation platform apertium.http://xixossna.dlsi.ua.es/fran/apertium2documentation.*
3. *http://linguistlist.org/issues/13/13-622.html*
4. *Ashwini Vaidya. 2009. Using paradigms for certain morphological phenomena in Marathi. In Proceedings of ICON.*
5. *http://aclweb.org/anthology/W12-2302*
6. *M. F. Porter. 1980. An algorithm for suffix stripping. Originally published in Program, 14 no. 3, pp 130-137.*
7. *R. Krovetz. 1993. Viewing morphology as an inference process. In Proceedings of COLING*