International Journal of scientific research and management (IJSRM)
\||Volume||3||Issue||6||Pages|| 3230-3233||2015||
Website: www.ijsrm.in ISSN (e): 2321-3418

Morphological Disambiguator for Marathi using NLP

Arti P. Khadtare[1], Dr. Suhas Raut[2], M. S. Otari[3]

[1]Department of Computer Science & Engineering,
Nagesh Karajagi Orchid College of Engg. & Technology,
Solapur, Maharashtra, India.

[2]Department of Computer Science & Engineering,
Nagesh Karajagi Orchid College of Engg. & Technology,
Solapur, Maharashtra, India.

[3]Department of Computer Science & Engineering,
Nagesh Karajagi Orchid College of Engg. & Technology,
Solapur, Maharashtra, India.

Abstract: *Natural Language Processing (NLP) is an emerging technology in recent years, which consist of Morphological analysis of languages. Morphological analysis gives description about words in that language. Morphological analysis of Indian languages is a difficult task. In an Indian language like Marathi word can have many inflections which results in ambiguity. So the description generated by analysis can be further used to find out the meaning of that word to remove ambiguity while translating from one language to another. This paper describes an approach for finding the meaning of the word, using description generated by the analysis and other information such as the position of the word in a sentence, suffix etc., to remove ambiguity when translating sentence written in Marathi (Devnagari) language to English language.*

Keywords: NLP, Transliteration, Devnagari Script, Morphological analysis.

1. Introduction

1.1 Morphological Analysis

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, stress, or implied context. Morphological analyzer is a system which takes a single word, at a time, as an input, performs its analysis with the help of a dictionary of root words, paradigm table and dictionary of indeclinable word and returns its description as a result. Fig 1.1 shows morphological analyzer. The linguistic resources required by the morphological analyzer include a lexicon and inflection rules for all paradigms. These are few linguistic resources [9].

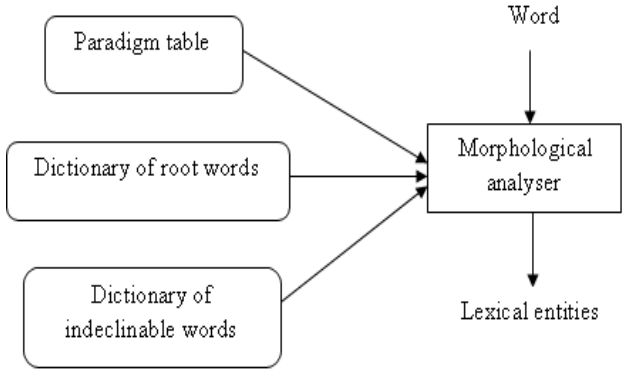

Fig.1.1: Morphological Analyzer

1.2 Marathi Language and need for Morphology

Marathi is the official language of the state of Maharashtra (India) and is one of the 23 official languages of India. There were 73 million speakers in 2001; Marathi ranks 20th in the list of most spoken languages in the world. Marathi has the fourth largest number of native speakers in India [10]. The goal of Natural Language Processing (NLP) is to build computational model of natural language for its analysis and generation, building intelligent computer system such as machine translation system, man-machine interfaces to computers in general, speech understanding system, text analysis and understanding system etc. [3]. Morphological analysis helps to do perform this kind of knowledge based extraction of languages.

1.3 Transliteration

Transliteration generation is a part of language processing. A transliteration generator is a program to generate the transliteration (representation in English characters) for an input word given in Marathi (Devnagari). In proposed system we have used the transliteration generator for Mangal font. In current techniques the transliteration generated for an input word depends on font used for typing the text. This transliteration of Marathi (Devnagari) words can be used for further text processing such as translation (i.e. translating text from one language into another), comprehension etc. or other NLP applications [11].

2. Methodology

2.1 General flow of Proposed System

- In the proposed system user will enter input sentence in Marathi (Devnagari script).
- Each alphabet of Marathi word, from the input sentence, will be transliterated, i.e. converted to equivalent English character, by transliteration generator.
- Then the sentence will be tokenized; sentence will be

divided into words.

- Now each word will be analyzed to find out description about word. Description about word will be found out with the help of paradigm table and dictionary of root words. [Algorithm: For performing word analysis].
- This information about word like type, gender, root etc. along with the information about suffix, affix and position of word will be used to determine appropriate meaning of word in the context of sentence.

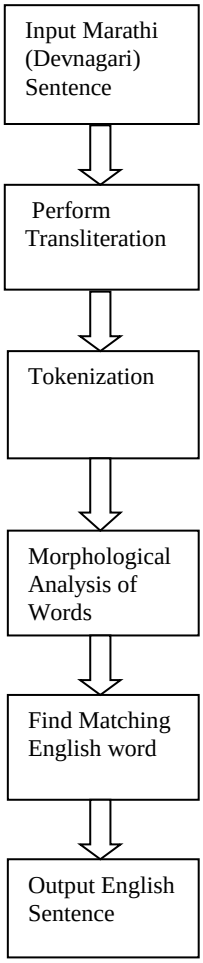

Figure 2.1 General flow of proposed system

- After this corresponding word will be replaced with its equivalent English word by searching in database. This whole process will be repeated for each and every word in the sentence. And then the English sentence equivalent to Marathi sentence will be displayed.
- Algorithm: For performing word analysis
  Purpose: To find root word, inflected form of word and its information from root word to which it belongs
  Input: Root Word table, Inflected word forms
  Output: Inflected word form with its root word and related information

  Algorithm:
  For every entry in root word table, do
  1. Retrieve original word root, org_wrd, and root word group, gr_wrd, to which given word belongs.
  2. Find all inflected form entries for that root word group, gr_wrd.
  3. For all entries (c (character/s), s (string of character)) of gr_wrd
     Add or delete 'c' from org_wrd
     Add 's' (if any) to org_wrd and form inflected word, infl_wrd
  4. Compare original word with infl_wrd
     If match found retrieve its related information and store it along with original root word
  5. If no match found
     Add default information to that word
  End algorithm
- For example if the given input sentence is 1. "पूजा शाळेत जाते." The input sentence will be transliterated as "pUjA shALet jAte". After tokenization the sentence will be divided into tokens: "pUjA", "shALet", and "jAte". Analyzing these tokens will provide description like,

original_word	root_word	type	gender
shALet	shALA	noun	feminine

In Marathi the word "पूजा" can be used as verb (which means worship in English) and as a proper noun (as a girl's name) and to determine the meaning of this word in the given sentence we are going to analyse the obtained description. From this description we can see that the word 'shALet' is an inflected form of root word 'shAlA'. Here the position of word in the sentence is second so we conclude that it is used as an object in the sentence and the suffix is 't' and requires "to" to be added as suffix in translated sentence. Similarly word at the third position in the sentence is 'jAte' which is inflected form of root word 'jA' which is a verb. Also suffix 'te' indicates it is in present tense. So we consider word 'pUjA' along with its first position as a subject in the given sentence and hence will be used as proper noun in translated sentence.

Now English equivalent words of these words (jA=go-present tense, shALa=school) will be replaced and translated sentence will be displayed as output. Hence the translated sentence will be "pUjA goes to school". Similarly in the second sentence, 2. "मी पूजा करतो.", "mI" is at first position and so it is considered as the subject here. "pUjA" is in second position and hence we consider it is used as a verb. And "krtao" is the helping verb in the sentence. Hence in this sentence "pUjA" is replaced as "worship". Consider two other sentences, e.g., 3. "माझे नाव राम आहे." and 4. "ही नाव मोठी आहे." In the given sentences the intended meaning of the word "नाव" is name (noun) and boat (noun) respectively. Here the meaning of the word will be determined with the help of the type and gender of first and second words of the sentence. In the first sentence the position of word "mAjhe" is first and it indicates possession and the word "rAm" (proper noun) is what is possessed. Also the gender of the world "mAjhe" is neutral. So here "nAv" is used as "name". In the second sentence the first word, "hI", is a pronoun and its gender is "feminine" also the third word, "maoThI", is an adjective but indicates it has "feminine" gender hence the gender of the word "nAv" should also match with these two and hence its meaning will be considered as "boat (noun, feminine)" in the given sentence. Like this by using more and different inflection rules and grammar rules, meaning of different ambiguous words in the context of sentence can be found out.

2.2 GUI Design

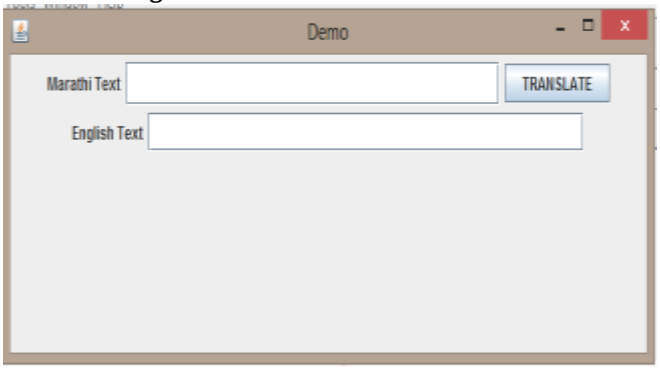

Fig. 2.2 GUI Design

In above GUI design, there are 2 text areas and 1 button.

- Text Area

1. First Text Area is for giving Input in Marathi Sentence in Devnagari script.
2. Second Text Area is for Displaying Output in English Text.

- Button

1. Translate Button: For performing analysis and displaying English equivalent sentence of given Marathi sentence.

3. Results

The following screen shots show the result of the proposed system. It consists of different examples explained above, in the section 2.1 general flow of proposed system, and the output of those sentences on google translator is shown for reference.

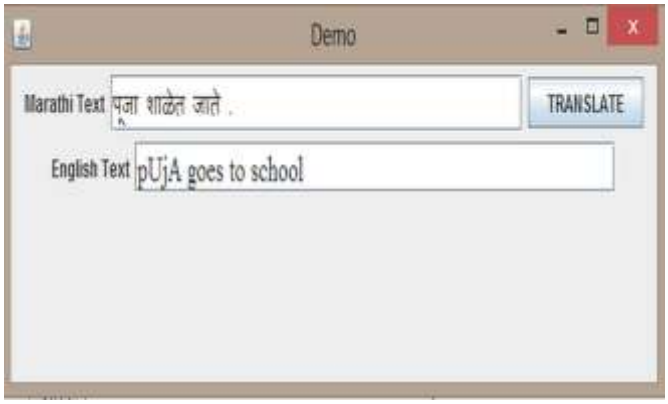

Fig 3.1 Input and Output of Example 1on proposed system

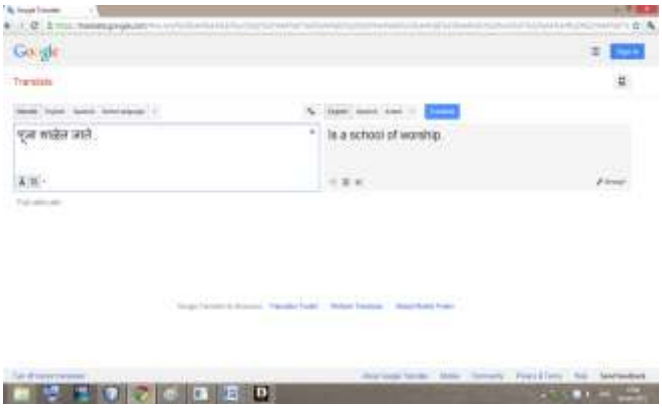

Fig 3.2 Output of Example 1 on google translator

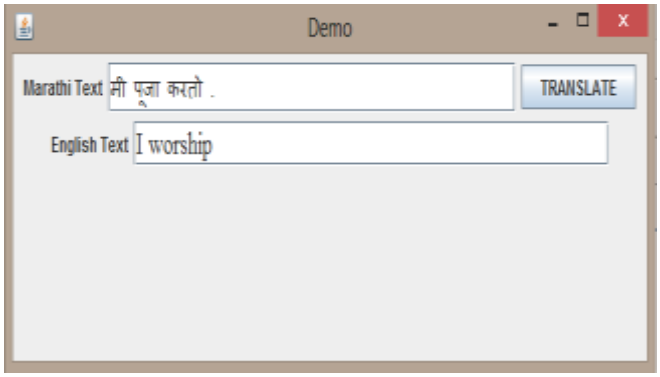

Fig 3.3 Input and output of Example 2 on proposed system

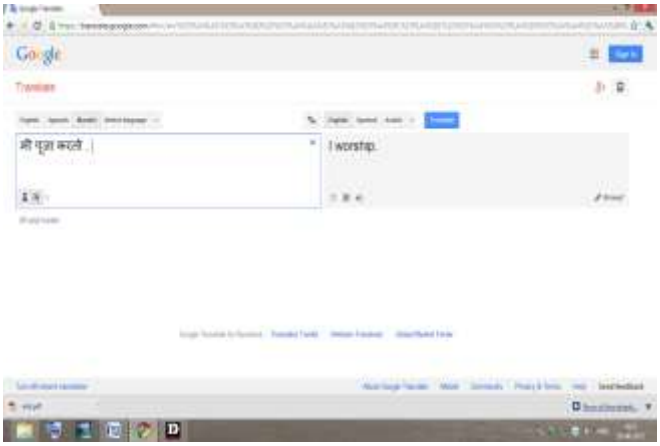

Fig 3.4 Output of Example 2 on google translator

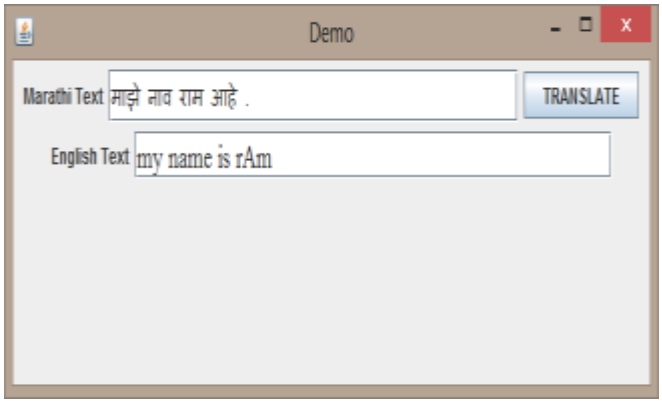

Fig 3.5 Output of Example 3on proposed system

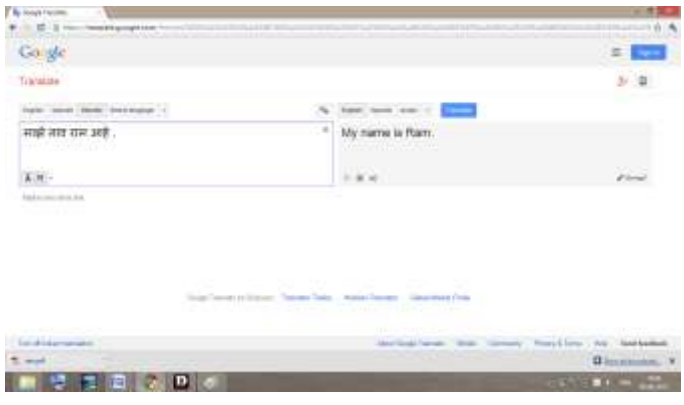

Fig.3.6 output of example 3 on google translator

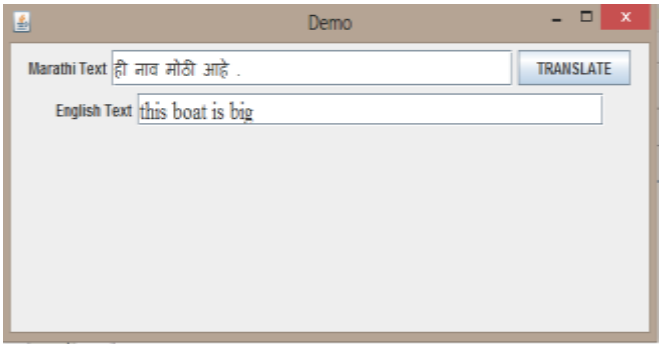

Fig. 3.7 Output of Example 4 on proposed system

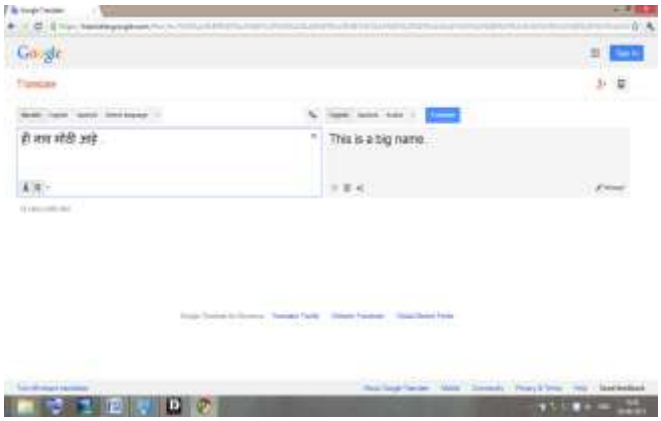

Fig. 3.8 Output of Example 4 on google translator

4. Conclusion and Future Scope

Thus this paper describes methodology of morphological disambiguator for Marathi. Proposed system will be able to find out the appropriate meaning of Marathi word in sentence depending upon the context. This will help in finding correct replacement for that word in English language to help improve translation of Marathi sentences to English sentences. Proposed system provides solution for simple and short sentences with ambiguous word; this can be improved for long and/or complex sentences by using more inflection and grammar rules. Also language sense methods, like lakshna chart, can be further used to remove ambiguity more accurately.

5. References

[1] A Paradigm-Based Finite State Morphological Analyzer for Marathi. Mugdha Bapat, Harshada Gune, Pushpak Bhattacharyya, 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.

[2] Morphological Analyzer for Marathi using NLP ,Pratiksha Gawade, Deepika Madhavi, Jayshree Gaikwad, Sharvari Jadhav, Rahul Ambekar, IJERA, ISSN: 2248-9622 Vol. 3, Issue 2, March-April 2013.

[3] Natural Language Processing: A Paninian Perspective. Bharati Akshar, Vineet Chaitanya, Rajeev Sanghal, Department of Computer Science and Engineering, Indian Institute of Technology Kanpur,1995.Publication Prentice-Hall of India, New Delhi.

[4] Plural Problems in the Nominal Morphology of Marathi Shalmalee Pitale, Vaijayanthi Sarma, 25th Pacific Asia Conference on Language, Information and Computation, pages 178–185, December 2011.

[5] Processing of Kridanta (Participle) in Marathi Ganesh Bhosale, Subodh Kembhavi, Archana Amberkar, Supriya Mhatre, Lata Popale, Pushpak Bhattacharyya, Proceedings of ICON-2011: 9th International Conference on Natural Language Processing Macmillan Publishers, India.

[6] Two-level Morphology: a general computational model for word-form recognition and production. Koskenniemi Kimmo, University of Helsinki, Department of General Linguistic, Hallituskatu 11-13, SF-00100, HELSINKI 10, FINLAND, PUBLICATIONS, No. 11, 1983.

[7] Using Paradigms for certain morphological phenomenon in Marathi. Ashwini Vaidya and Dipti Misra Sharma, Proceedings of ICON-2009: 7th International Conference on Natural Language Processing, Macmillan Publishers, India.

[8] Verbs are where all the action lies: Experiences of Shallow Parsing of a Morphologically Rich Language. Harshada Gune, Mugdha Bapat, Mitesh M. Khapra, and Pushpak Bhattacharyya, Proceedings of COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics: Pages 347-355, Association for Computational Linguistics Stroudsburg, PA, USA 2010.

[9] Wikipedia-https://en.wikipedia.org/wiki/Morphology _ (linguistics)

[10] Wikipedia-https://en.wikipedia.org/wiki/Marathi_ language.

[11] Transliteration Generator for Devnagari using Mangal Font, IJAIEM, Volume3, Issue8, 2014.