Morphological Analyzer for Gujarati using Paradigm based approach with Knowledge based and Statistical Methods

Jatayu Baxi
Computer Engineering Department
Dharmsinh Desai University
Nadiad
jatayubaxi.ce@ddu.ac.in

Pooja Patel
Computer Engineering Department
Dharmsinh Desai University
Nadiad
poopatel2912@gmail.com

Brijesh Bhatt
Computer Engineering Department
Dharmsinh Desai University
Nadiad
brij.ce@ddu.ac.in

Abstract

Morphological Analyzer is a tool which performs syntactic analysis of a word and finds root form of input inflected word form. Morph analyzer serves as a pre-processing tool for many NLP applications. Significant amount of work has been done in this area for many Indian languages but not much work has been reported for Gujarati language. We present Morph analyzer for Gujarati language. The Morph analyzer is developed using a hybrid approach that combines statistical, knowledge based and paradigm based approach. We present detailed study of different approaches. We demonstrate a significant improvement in overall accuracy and achieve 92.34% and 82.84% accuracy with knowledge based hybrid method and statistical hybrid method respectively.

1 Introduction to Morph Analyzer

Morphology is a branch of linguistics that carries out study of words, their internal structure and their meanings. A morpheme is the smallest grammatical unit in a language.

Developing an accurate Morph analyzer is a challenging task, particularly for highly inflectional and agglutinative language. In order to develop Morph analyzer for Gujarati, we studied inflections of Gujarati language and identified various grammatical paradigms. We have identified 8 paradigms for Noun , 3 paradigms for verb and 3 paradigms for Adjectives. Paradigm based approach suffers when a word falls into more than one paradigm because of common inflections. We reduce this problem by consulting knowledge base ‘wordnet’ and by performing corpus based statistical analysis of a word.

The rest of the paper is organized as follows: Section 2 discusses survey and comparison of existing approaches. Section 3 describes paradigm construction process and various paradigms. Section 4 discusses hybrid method used to build Morph analyzer for Gujarati and section 5 shows experiment and evaluation.

2 Related Work

Morphological analyzer has been a continuously evolving area of research. A lot of work has been done for English language. However it remains challenging task to develop Morph analyzer for highly agglutinative and inflectional languages

The first Morphological analyzer system is KIMMO(Karttunen, Lauri, 1983) which follows two level morphology approach. This approach is suitable for languages with less degree of inflection. For higher degree of inflection, the model does not perform well. Some unsupervised methods for developing Morph analyzer have also been tried. Hammarstorm and Borin,2003 presented survey based on various Unsupervised Learning Techniques for Morphological analyzer. The input to such algorithm is raw Natural Language text data. This approach requires large training corpus

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 178–182,
Trivandrum, India. December 2015. ©2015 NLP Association of India (NLPAI)

and hence in absence of sufficient corpus, the machine learning efficiency would be less. Niraj Aswani et al,2010 used unsupervised Method that takes both prefixes as well as suffixes into account. Given a corpus and a dictionary, this method can be used to obtain a set of suffix-replacement rules for deriving an inflected word's root form. This system is basically built for Hindi but some experiments are done on Gujarati language also. Akshar Bharti et al, 2001 presented an algorithm for unsupervised learning of morphological analysis and morphological generation for inflectionally rich languages. The method depends on variety of word forms present in the existing corpus.

Beesley, 2003 proposed concept of Finite state Morphology. and this approach was used to build Morphological analyzers for high inflectional languages, as language can be easily represented by Finite state machine. In this approach, XFST is used as an interface. It is an interface which gives access to finite state operations. The interface of XFST includes a lookup operation and generation operation. Hence this approach can be used for Generation and Analysis of Morphology.

Considerable amount of work has been done for the languages such as Hindi,Marathi,Tamil,Malayalam,Bangla also. Based on FST approach proposed by Bessley, (Akshar Bharti) proposed paradigm based approach for building Morphological analyzer. In this approach, the language expert provides different tables of word forms covering the words in the language. Set of roots covered by particular table have similar inflectional behaviour. This approach also requires dictionary with root of the word and the paradigm to which the root belongs. (Jyoti Pawar et al, 2012) discusses work in which the morphological analyzer for Konkani has been developed using FSA based approach with Word paradigm model. Unlike traditional method, they have sequenced morphemes. (Harshada Gune et al, 2010) developed paradigm based Finite state Morphological analyzer for Marathi.

(Rajendra Rajeev, 2011) developed Morphological analyzer using suffix stripping approach for Malayalam language. The finite state transducer is used to sequence the morphemes and to validate the ordering. In this system the Suffix stripping method with sandhi rules are used that does not require any lookup tables.

As far as work for Gujarati language is concerned (kashyap Popat et al, 2010) have developed lightweight stemmer for Gujarati language using hand crafted suffix rules. List of hand crafted Gujarati suffixes which contains the postpositions and the inflectional suffixes for nouns, adjectives and verbs are created for use in this approach.

(kartik Suba et al, 2011) developed Inflectional and derivational stemmer. The inflectional stemmer is built using hybrid approach and derivational stemmer is built using rule based approach. Four lists of suffixes which contain postpositions and inflectional suffixes respectively for nouns, verbs, adjectives and adverbs are created.

3 Paradigm Construction

Gujarati is morphologically rich language. Based on various grammatical features, single root may generate multiple word forms Words can be categorized into paradigms based on the similarity in grammatical features and word formation process. It is observed that for a particular grammatical feature there are some similarities in word formation process.

A Paradigm defines all word forms that can be generated from given stem along with grammatical feature set associated with each word form. For paradigm construction, sample corpus of inflected words is taken. A list is prepared for all possible suffixes from the sample data. The words which take similar set of suffixes are grouped into single paradigm. For example word સ્ત્રી(Stri) and છત્રી(chaatri) both take same inflection for plural transformation (સ્ત્રીઓ and છત્રીઓ) so we group them under single paradigm. For all the words belonging to same paradigm, the rule to form root word and set of possible suffixes to generate other word forms remains same.

Gujarati nouns inflect in gender and number and case. Gujarati has three genders and two numbers. Table 1 shows 7 paradigms identified for Gujarati noun. The first column shows paradigm ID, Second column shows suffix list which is used to detect this paradigm, third column shows one inflected word belonging to that paradigm and final column shows corresponding root word. We observe from table 1 that same suffix may belong to more than one paradigm.

Id	Suffix	Example	Root Word
1	◌ો,◌ી,◌ુ,◌ા ◌ાઓ,◌ીઓ	છોકરાઓ (Chokrao)	છોકરું (Chokru)
2	◌ો,◌ી,◌ા	▯દિકરો (Dikro)	દિકરો (Dikro)
3	◌ો,◌ી,◌ુ,◌ા	કટકો (Katko)	કટકો (Katko)
4	◌ો	વાક્યો (Vaakyo)	વાક્ય (Vaaky)
5	◌ી,◌ીઓ	સ્ત્રીઓ (Strio)	સ્ત્રી(Stri)
6	◌ો,◌ા	મહિના (Mahina)	મહિનો (Mahino)
7	◌ો,◌ા	એકતા (Ekta)	એકતા (Ekta)

Table 1: Noun Paradigms

Id	Suffix	Example	Root Word
1	◌ો,◌ી,◌ુ	સારો (Saro)	સારું (Saru)
2	No Inflection	સરસ (Saras)	સરસ (Saras)
3	◌ી	વ્યભિચારી (Vyabhichari)	વ્યભિચારી (Vyabhichari)

Table 2: Adjective Paradigms

Table 2 shows various paradigms for adjectives. Adjectives can be classified into variant and non-variant adjectives based on the inflections that they take. A non-variant adjective do not inflect with gender.

Gujarati verb inflect in gender, number, case and tense. Verb and adjective require gender agreement with Noun. Table 3 shows various paradigms for verb. We observe that unlike Hindi, Gujarati verbs inflect in gender for only past tense. For present and future tense, verb does not inflect with gender

4 Implementation

In this section we describe paradigm based, knowledge based and statistical approach. We also propose ways to merge these approaches.

Id	Suffix	Example	Root Word
1	આડે, આડી, આડશે, અડ્યા, એ, ઈશ, જો, ઓ, વું, તો, તી.	જમાડ્યું (Jamadyu)	જમ્યું (Jamvu)
2	આવવું, વ્યું, આવશે, આવ્યા, એ, ઈશ, જો, ઓ, વું, તો, તી.	કરશે (Karshe)	કર્યું (Karvu)
3	ધું, વ્યું, આવશે, આવ્યા, ય, જો, વ, વું, તો, તી, ◌ુ	ખાધું (Khadhu)	ખાવું (Khavu)

Table 3: Verb Paradigms

4.1 Paradigm Based Method

We define paradigms for various part of speech for Gujarati language. Paradigm building methodology is covered in the next section in brief. Each paradigm consists of rules to obtain root word, various inflections possible with that paradigm and a representing word for that paradigm. An inflected word is given as an input to the system. The system checks suffix and determines one or more matching paradigm for input word. It is possible that the same rule is present in more than one paradigm. For each matched paradigm system applies rule to generate root word and gives root word as output.

It is possible that a single inflected word is mapped with more than one root word. So to eliminate this multiple output we incorporate two different methods knowledge based and statistical method.

4.2 Knowledge Based Method

Whenever an inflected word gives more than one Root word as an output using paradigm based method, We consult Gujarati Wordnet knowledge base and select best output. Gujarati wordnet contains around 81000 words in root form. Out of all multiple outputs, correct output is matched with wordnet and it is chosen as correct output. The success of this approach lies in the coverage of various root words in the dictionary.

4.3 Statistical Method

Statistical approach focuses on disambiguating multiple outputs based on heuristic developed on length of an inflected word and transformation process for generating root word. In this method we construct probability table which contains transformation rules along with its probability which is determined using set of training words. Snapshot of this table is shown in Table 4. For example if an inflected word ends with ી then according to Table 4 it has two possible transformations ી - > ુ + ં and ી -> ી. Training data suggests that probability of rule 2 is higher than rule 1 so in this case transformation 1 would be selected for root word generation.

Sr No	Suffix	Possible Transformation	Probability (%)
1	ો	ો - > ો	90.3
		ો- > ુ + ં	32.6
		ો- >No Suffix	2.36
2	ી	ી- > ી	82.36
		ી- > ુ + ં	12.28

Table 4: Statistical Method

5 Experiment Setup

For our experiment we have prepared Gold set of data which consists of 200 Nouns, 200 verbs and 100 Adjectives. Input to Morph analyzer is inflected word and Output is root word. We have root word information about this Gold data which is manually prepared by Linguists. We compare the result produce by our system with actual root word in the Gold data. Analysis of the result is based on factors like How many words are mapped to correct paradigm, How many words are not mapped to correct paradigm , How Many Words belongs to Multiple Paradigm, which category gives good result in the system. From these parameters we analyze accuracy of our system and analyze how various part of speech affects accuracy of the system.

We classify outputs into 2 different categories:

- Correct: Words for which only one output is generated which is correct output.
- Incorrect: Words for which one output is generated and which is incorrect.

Due to the fact that same rule can be part of multiple paradigms, for many words we get multiple outputs. For all those words we apply knowledge based and statistical methods for output enhancements and evaluate accuracy of the system after applying each of the above method.

6 Result Analysis

Simple rule based method has problem of multiple paradigm output. For the words giving multiple outputs we apply knowledge based method and statistical method for removing incorrect outputs. Table 5 and Table 6 summarize results obtained after applying knowledge based and statistical methods. Table 7 shows accuracy of both knowledge based and statistical methods. It can be seen that problem of multiple outputs in case of simple rule based technique can be handled by knowledge based and statistical methods.

POS	Correct	Incorrect	Accuracy
Noun	197	3	98.5
Verb	179	21	89.5
Adjective	89	11	89

Table 5: Results: Knowledge Based Method

POS	Correct	Incorrect	Accuracy
Noun	154	46	77
Verb	149	51	74.5
Adjective	97	3	97

Table 6: Results: Statistical Method

Following points were observed as a part of error analysis.

- For Rule based approach, same suffix rule may apply to different categories so number of words correctly classified is less and degree of multiple outputs is more.
- For Knowledge based method, the reason for error is that some words may not be present in wordnet dictionary so knowledge based approach fails to remove ambiguity from such words.

Approach	Accuracy (%)
Rule Based	58.26
Knowledge Based	92.34
Statistical	82.84

Table 7: Approach wise Accuracy

- Statistical Methods do not depend on any dictionary but they are probabilistic. So depend upon training data used, accuracy may vary.

7 Conclusion and Future Scope

In this paper we presented a Gujarati morphological analyzer. We choose paradigm based approach for our implementation and analyze results obtained for gold set of words. We incorporate two output enhancement methods knowledge based method and statistical methods. We prepared gold set of words and tested them with our system and analyzed results. We conclude that after removing multiple outputs system gives average accuracy of 92.34 for knowledge based method and 82.84 for statistical method. The limitation of current system is that it cannot handle derivational morphology.

References

M.F.Porter 1972. An algorithm for suffix stripping, volume 1. Program,14:130-137

Gulsen Eryigit and Adali Esref 2004. An affix stripping morphological analyzer for turkish, volume 1. IASTED International Multi- Conference on Artificial Intelligence and Applications,pages 299–304, Innsbruck, Austria.

Lauri Karttunen and Kenneth R. Beesley 2003. Finite State Morphology

Akshar Bharti Natural Language Processing - A Paninian Perspective

Harald Hammarstorm and Lars Borin 2004. Unsupervised learning of morphology, volume 1. Computational Linguistics June, 37:309–350, June 2011.

Jyoti Pawar Shilpa Desai and Pushpak Bhattacharya 2012. Automated paradigm selec-tion for fsa based konkani verb morphological analyzer, COLING 2012, Mumbai,India, 10-14 Dec, 2012.

Harshada Gune Mugdha Bapat and Pushpak Bhattacharyya 2010. A paradigm-based finite state morphological analyzer for marath, Workshop on South Asian and South East Asian NLP (part of COLING 2010), Beijing, China, August 2010.

Kashyap Popat Pratik Patel and Pushpak Bhattacharyya 2010. Hybrid stemmer for gujarati, International Conference on Computational Linguistics (COLING), pages 51–55, Beiging, August 2010.

Dipti Jiandani Kartik Suba and Pushpak Bhattacharyya 2011. Hybrid inflectional stemmer and rule-based derivational stemmer for gujarati, IJCNLP 2011 Workshop on Segmentation and Morphology for South Asian Languages, Chiang Mai,Thailand, November 2011.

Niraj Aswani and Robert Gaizauskas 2010. eveloping morphological analysers for south asian languages: Experimenting with the hindi and gujarati languages, Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10), Valletta,

Rajendran N Rajeev R and Elizabeth Sherly 2011. A suffix stripping based morph anal-yser for malayalam language, Dravidian Studies a Research Journal, pages 61–72.

George Cardona 1965. A Gujarati Reference Grammar

Colin Masica 1991. The Indo-Aryan Languages Malta, may 2010. European Language Resources Association (ELRA). ISBN 2-9517408-6-7.

W.S. Tisdall 1892. A Simplified Grammar of the Gujarati Language

Brijesh Bhatt, Dinesh Chauhan, Pushpak Bhattacharyya, C.K. Bhensdadia and Kirit Patel. 2012. Introduction to Gujarati Wordnet International Conference on Global Wordnets (GWC 2011),Matsue,Japan.

Karttunen, Lauri. 1983. KIMMO: A general morphological processor In: Linguistic Forum 22, (1983) 163–186

Nikhil Kanuparthi,Abhilash Inumella and Dipti Misra Sharma. 2012. Hindi Derivational Morphological Analyzer Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology (SIGMORPHON2012) , pages 10–16

Akshar Bharti, Rajeev Sangal, S.M.Bendre, Pavan Kumar, Aishwarya 2001. Unsupervised improvement of Morphological Analyzer for inflectionally Rich Languages Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium