Rule Based Gujarati Morphological Analyzer

Utkarsh Kapadia[1] and Apurva Desai[2]

[1] **Department of Computer Science, Veer Narmad South Gujarat University**
Surat, Gujarat 395007, India

[2] **Department of Computer Science, Veer Narmad South Gujarat University**
Surat, Gujarat 395007, India

Abstract

Gujarati is an Indian Language spoken widely by over 50 million people of Gujarat in India and abroad. Gujarati like other Indo-Aryan languages like Hindi, Marathi is morphologically rich. Morphological analysis is an important step for many Natural Language Preprocessing (NLP) applications like machine translation, grammar inference, and information retrieval etc. In this paper we have presented morphological analyzer on rule based approach. Lexical dictionary of root words is created. Manually crafted rules with linguist are developed. The analyzer tool takes Gujarati sentence as an input, and produces its grammar class, gender, number, and tense and person information with its root words. The tool works on both inflectional and derivational morphemes. We have obtained accuracy of 87.48% upon evaluation with text taken from essays and short stories.

Keywords: *Gujarati, Morphological Analyzer, Rule based, Natural language Processing, Part of Speech Tagging.*

1. Introduction

Morphological analysis is identifying root form of word and producing grammar class with person, gender, and number information. Morpheme is the smallest grammatical unit of natural language. Each word is comprised of one or more morphemes. Morphology can be categorized in to two types: inflectional and derivational. In inflectional morphology word does not change its grammatical class when combined with morpheme while in derivational it results in different class as well meaning. Morphemes can be also classified as either free morphemes or bound morphemes. Free morphemes can appear independently in sentence while bound morpheme can only appear with other free morphemes to form a word.

Considerable amount of work has been done in area of morphological analyzer and stemmer of natural languages. There are two types of approaches that are found in litterateurs namely supervised or semi-supervised and unsupervised. First approach uses hand-coded suffix replacement rules and lexicon for stemming while in second approach, rules are derived from corpus automatically. First approach being language specific requires considerable linguistic expertise to craft rules, but it can result in higher performance [3]. In second approach, rules are derived from corpus automatically. Morphological analyzer and generator work for Hindi was carried out by Vishal G. & Lehal G.S [1]. Their work mainly focuses on inflectional morphology. They mentioned that most of Hindi nouns inflections can take up to 8 forms and verbs can take up to 50 forms. They created a list of paradigms that is followed by a group of words. They also stored all commonly used word forms in database but they excluded proper nouns. They claim that the approach prefers time and accuracy over space. Niraj A & Robert [2] extended wordlist of Shrivastava[3] by adding those words which were there in EMILLE corpus but not in the wordlist based on suffix analysis. Their rules were derived automatically from corpus and dictionary by replacing one character at a time from right and matching resulting form with root list. If suffix is found, rule is formed. Then they computed probability of suffix based on count of suffix appearing in corpus. Subsequently rules were applied with priority and length of suffix. Priority was based on probability of suffix appearing in corpus. They have reported Precision=0.821, Recall=0.803 and F Score=0.812 with extended WorldNet and rule set. Baxi & others [5] demonstrated paradigm based approach combined with statistical approach and reported accuracy of 82.84%. Finite State [6,7] morphological analyzer is also demonstrated for Marathi and Hindi with accuracy in Marathi of 97% and that of Hindi was 93%. Acquisition of morphology from corpus using unsupervised approach for Assamese was demonstrated by Utpal & Others [8]. In their work they mentioned that suffix list and lexicon can improve overall accuracy of the system. Nikhil & others [9] produced derivational morphological analyzer based on inflectional analyzer produced by IIT Hyderabad. They did manual process of obtaining derivational suffixes of Hindi and obtained 22 suffixes and rules. They were able to improve overall inflectional analyzer accuracy by 5%.

2. Gujarati Lexicon Preparation

There are fifty letters in Gujarati alphabet – sixteen vowels, and thirty four consonants according to Devanagari characters, but only 11 vowels and 29 consonants are used commonly. The words of Gujarati are arranged under five classes, called Parts of Speech. The names of these parts of Speech are: Noun, Pronoun, Adjective, Verb, other words. Noun admits of inflection to express Number, Gender and case. There are two numbers, the singular and the plural. There are three genders: masculine, feminine and neuter. Cases in Gujarati are seven omitting vocative. They are nominative, agentive, accusative/dative, genitive, instrumental and locative.

2.1 Nouns

Most Gujarati nouns are ending in vowels e.g. અ, આ, ઇ, ઉ, એ, ઓ, ઐ etc. While less nouns ending in consonants e.g. ખ, ઠ, શ. Gujarati nouns are formed by: Noun stem + Gender Marker + Number Marker + Case Marker [4]. E.g. છોકરાઓને (boys) can be expressed by: છોકર + ◌ા + ઓ + ને. Unlike Gujarati, Hindi Case markers are written separately from word e.g. लडको ने. Morphological analysis of Gujarati shall be different from language like Hindi even both of them belongs to same Indo-Aryan family

Root noun forms listed with class, number and gender information. There are 13,964 nouns tagged with gender and number information. Sample of such nouns are listed in table 1.

Table 1: Noun List

Word	***Tag***	***Number***	***Translation***
અક્કલ /akkala/	NNF	S	intelligence
અકળામણ/akaḷāmaṇa/	NNF	S	anxiety
અખરોટ /akharōṭa/	NNN	S	walnut
અગત્યતા /agatyatā/	NNF	S	importance

2.2 Pronouns

Gujarati pronouns decline with persons (first, second and third), numbers (singular, plural) and cases. They have also inclusive and exclusive contrast in third person plural. In addition, their second person plural form is also used as honorific. Pronoun being closed class, a list of 238 pronouns prepared in various sub categories like personal, demonstrative, interrogative, relative, reflexive, reciprocal and indefinite

2.3 Adjectives

In Gujarati, adjectives precede the nouns which they qualify. Adjectives are of two types: declinable (vikārī) and Indeclinable (avikārī). Variable (declinable) adjectives vary in terms of the gender and number of the nouns they modify, whereas the invariable adjectives do not vary. According to grammar they can be further classified in adjective of quantity, quality, number, demonstrative, and interrogative etc. There are currently 3892 adjectives in lexical database. Sample adjectives are listed in table 2.

Table 2: Adjective List

Word	***Tag***	***Translation***
અકબંધ /akabandha/	JJ	intact
અકળ /akaḷa/	JJ	weird
અખૂટ /akhūṭa/	JJ	inexhaustible
અખિલ /akhila/	JJ	whole

2.4 Verb

Gujarati verbs (non-inflected) have the following structure: verb stem + inflectional material. Inflectional material may consists of various features such as tense, person, gender. Sample list of verb and its tag are shown in table 3. There are 1056 distinct verbs base forms present in lexicon database

Table 3: Verb List

Word	***Tag***	***Translation***
અચકાવું/ acakāvuṁ/	VM	hesitate
અજમાવવું/ ajamāvavuṁ/	VM	try
અજવાળવું/ ajavāḷavuṁ/	VM	illuminate

2.5 Other words

Gujarati language has other words like post-positions, connections, interjections, negations, compound words etc.

In derivational morphology, word class is changed when suffix is attached to stem. There are such 22 suffixes separated to identify derived nouns. E.g. કર (do) + નાર = કરનાર (doer). Such nouns are formed by suffix attachment with either adjectives or verbs or even noun, which results in change of meaning or grammar class.

Complete database statistics is given in table 4.

Table 4: Word Database Statistics

Class	***Entries***
Adjectives	3892
Adverbs	172
Verb	1056
Noun	13964
Proper Nouns	8495
Pronouns	238
Others	314
Total	***28131***

3. Gujarati Morphological Formations

Rules for replacements are divided into three categories, noun, verb inflectional rules and derivational morphological rules. Noun rules are divided into case marker, number and gender marker replacement rules.

3.1 Noun Inflection Rules

Gujarati words appear in sentence with case marker which is to be stripped off before any further analysis. So for the reason we have found that we have to assign simple priority to rules to find stem from inflected or derived word. Such 12 suffixes replacement rules are separated. Some of the case marker rules are listed in table 5.

Table 5: Case Marker Replacement

Affix	***Replace***	***Order***	***Position***	***Example***
એ	-	1	1	છોકરાએ→છોકરા
ને	-	1	1	છોકરાને→છોકરા

Second replacement rules are number marker replacement rules after case marker replacement. These rules help in conversion of plural nouns to singular nouns. Some of these types of replacement rules are listed in table 6.

Table 6: Number Marker Replacement

Affix	***Replace***	***Order***	***Position***	***Example***
ા	ું	2	1	ગામડા→ ગામડું
ઓ	-	2	1	છોકરાઓ→છોકરા

Gujarati nouns also admit inflections as per three genders masculine, feminine and neuter. Rules formed helps to find word stem for noun category. Table 7 lists some of the rules for gender inflection.

Table 7: Gender Marker Replacement

Affix	***Replace***	***Order***	***Gender***	***Position***	***Example***
ો	-	3	1	M	છોકરો→ છોકર
ી	-	3	1	F	છોકરી→છોકર

3.2 Verb Inflection Rules

Gujarati verbs admit inflections as per gender, number, person, tense, aspect etc. Presently we have rule file with 65 verb replacement rules. Table 8 lists some of the rules for Gujarati verb.

Table 8: Verb Inflection Rules

Affix	***Replace***	***Order***	***Gender***	***Position***	***Example***
ીશ	વું	4	1	Fut.	રમીશ→ રમવું
્યો	વું	4	1	Past	રમ્યો→ રમવું
ું	વું	4	1	Present	રમું→ રમવું

3.3 Derivational Morphology

Gujarati language nouns can be formed by adding derivative suffix either from noun, adjectives or even verbs. There are 22 such commonly used noun endings identified. Some of them are listed in table 9.

Table 9: Derivational Morphology Rules

Affix	***Replace***	***Order***	***Class***	***Example***
નાર	-	5	Noun	રમનાર→રમ
ખોર	-	5	Noun	બડાઈખોર→બડાઈ
ગણું	-	5	Noun	પાંચગણું→પાંચ

All rules are grouped as per order of application on word. There are total 168 rules present in database.

4. System Description

4.1 Analyzer Algorithm

Firstly, we performed stemming guided by rules of language morphology which is about formation of admissible words. Morphemes are smallest unit of language and they carry some grammatical meaning. So morphemes should be separated linguistically.

For each word following steps are performed:

Step1: Word is searched against all possible roots in present in database of all grammar class to make sure if word is in the root form. Such roots are listed in table 1, table 2 and table 3. If found produce the class else go to Step2.

Step2: Word is matched with all case marker replacement rules suffix of table 5, if appropriate match is found it is replaced with replacement. Go to Step3.

Step3: Noun Analysis

Step3.1: Word is searched against root forms of noun class to check if word is noun root form of table 1, if found then grammar class information is displayed if not found then go to 3.2.

Step3.2: Word is searched against noun number marker replacement rules of table 6, replacement will occur if matching suffix is found and perform search of 3.1 If not found then Go to Step3.3.

Step3.3: Word is searched against gender marker rules, of table 7, replacement will occur if matching suffix is found and perform search of Step 3.1 if not found go to Step4.

Step4: Verb Analysis

Step4.1: Word is searched against inflection rules presented in table 8. Replacement will occur for matching suffix.

Step4.2: Check for Verb root in verb root table 3. If found its class information is presented. Go to Step 5

Step5: If word is not inflected then it is searched against in table 9 for derivational suffix and class information is presented if suffix matches.

4.2 Algorithm Analysis

Consider following cases:

Case I: **Input word** = છોકરીઓના
Case Marker = ના → છોકરીઓ
Number Marker = ઓ→છોકરી
Fem. Gender marker = ◌ી → છોકર
Suffix =◌ીઓના
Stem = છોકર
Result1: Category = NNF.PL.GEN.
Verb Rule: NULL
Result2: Verb not found

Case II: **Input word** = અધિકારીઓએ
Case Marker = એ→ અધિકારીઓ
Number Marker = ઓ→અધિકારી
Fem. Gender marker = ◌ી → અધિકાર
Suffix =◌ીઓએ
Stem = અધિકાર
Result1: Category = NNM.PL.ACCU.
Verb Rule: NULL
Result2: Verb not found

In Case II, feminine gender marker is present in the word but final category is masculine, as algorithm searches before replacing gender marker suffix, search which will lead to produce correct result.

Case III: **Input word** = કવીતા
Result1: Category = NNF.SG.
Verb Rule: તા → કવીવું
Result2: Verb not found

In Case III, although verb suffix is present, but after replacement word is not found in the verb list, so final category is identified as noun.

Case IV: **Input word** = રમીશ
Case Marker = NULL
Number Marker = NULL
Fem. Gender marker = NULL
Result1: Noun not found
Verb Rule: ◌ીશ → રમવું
Stem = રમ
Result2: Category=VM.FUT.SG

Gujarati verb is analyzed first against noun inflection and then against verb suffixes as shown above.

Case V: **Input word** = રમતો
Case Marker = NULL
Number Marker = ◌ો → રમત
Gender marker = NULL
Result1: Category: NNF.PL.
Verb Rule: તો → રમવું
Stem = રમ
Result2: Category: VM.SPST.SG.

In Case V, input word રમતો has two meaning they are: games (noun) and played (verb) both are from different grammar class. Due to this ambiguity, analyzer algorithm produces both of the possible grammar classes in results.

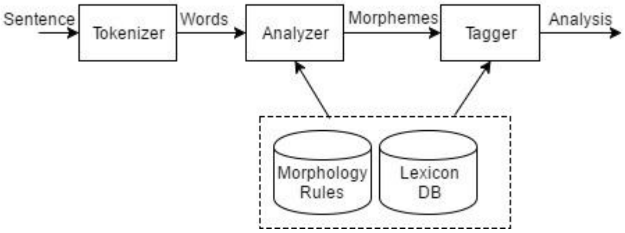

Fig. 1 Proposed Morphological Analyzer

https://doi.org/10.20943/01201702.3035

4.2 POS Tagger

Table 10 shows output of Part-of-Speech (POS) tagger.

Table 10: Results of Sentence Tagging

Sentence Input	***Tagged Output***
તેણે ઊંડો શ્વાસ લીધો.	તેણે/PP ઊંડો/JJ શ્વાસ/NNM લીધો/VM ./PUNC
અમે આ સમાચાર નથી સાંભળ્યા.	અમે/PP આ/DP સમાચાર/NNM નથી/NEG સાંભળ્યા/VM ./PUNC

5. Evaluation & Results

In experiment we evaluated performance of system in terms of accuracy, recall, precision and F-score.

Precision is defined as ratio of number of correct POS tags assigned by system to that of number of tags assigned by system. Accuracy is termed as fraction of words tagged correctly. Mathematically, precision is calculated using below formulae:

$$Precision\ (P) = \frac{No\ of\ Correct\ POS\ tags\ assigned}{No\ of\ tags\ assigned} \quad (1)$$

Recall is defined as ratio of number of correct POS tags and number of total POS tags in the text.

$$Recall\ (R) = \frac{No\ of\ Correct\ POS\ tags\ assigned}{No\ of\ POS\ tags\ in\ text} \quad (2)$$

F-score is calculated as:

$$F - score = \frac{(2 * Precision(P) * Recall(R))}{(Precision(P) + Recall(R))} \quad (3)$$

In order to evaluate the performance of proposed morphological analyzer, we developed corpus of 500 sentences (2514 words) of short stories taken from various Gujarati websites and stories article from corpus EMILLE [11]. The overall accuracy achieved by system is 87.48%. Following table11 shows test result of proposed system on text from two different domains.

Table 11: Result of Sentence Tagging

Test	***Precision***	***Recall***	***F-Score***
Essays	0.843	0.652	0.735
Stories	0.875	0.662	0.753

6. Conclusion & Future Scope

We have developed morphological analyzer for Gujarati based on hand coded inflectional and derivational rules obtained from regularities of words and affixes. The same was also used in POS tagger for sentence tagging. We have also given detailed description of morphological phenomena present in Gujarati. Apart from high accuracy, there are certain limitations of rule based system. First, it is difficult to build exhaustive rule based system for any natural language. Second, system may not produce any result if a suffix does not match with any of the rule. Since rules are highly dependent on language, it is challenging to make it language independent.

Appendix A

Standard POS Tag set

Sr	***Type***	***Sub Type***	***Label***	***Example***
1	Noun	Common	NNM	ચશ્માં
			NNF	ચોપડી
			NNN	પુસ્તક
		Proper	NNP	મોહન,રવિ, રશ્મિ
2	Pronoun	Personal	PRP	હું, તું, અમે
		Relative	PRL	જે,જેણે,જયારે,જ્યાં
		Reflexive	PRF	પોતે, સ્વયં
		Reciprocal	PRC	પરસ્પર
		Interrogative	PRQ	ક્યાં,ક્યારે
		Indefinite	PRI	કોઈ, કંઈ
3	Demonstrative	Deictic	DMD	ત્યાં, અહીંયા
		Relative	DMR	જેણે, જે
		Wh-word	DMQ	કોણ
		Indefinite	DMI	કોઈ
4	Verb	Main	VM	રમ, જમ, હસ
		Auxiliary	VAUX	છું, છીએ, છે
5	Adjective		JJ	સુંદર, સારું,ખરાબ
6	Adverb		RB	જલ્દી,ફટાફટ
7	Postpositions		PSP	સાથે

8	Conjunctions	Coordinator	CCD	અને, અથવા
		Subordinator	CCS	કારણકે
9	Particles			
		Default	RPD	પણ
		Interjection	INJ	જ, માત્ર
		Intensifier	INTF	બહુ, ઘણું
		Negation	NEG	નહિ, ના
10	Quantifier	General	QTF	થોડુંક, વધારે
		Cardinal	QTC	એક, બે, ત્રણ
		Ordinal	QTO	પહેલો, બીજો
11	Residuals			
		Foreign	RDF	ટીવી, કોર્ટ
		Symbol	SYM	$,*,&
		Punctuation	PUNC	. , ; : {} ( )
		Unknown	UNK	
		Echo words	ECH	કામ-બામ

Acknowledgments

UGC India has sponsored this project under SAP – II (file number is F.3-48/2011) and funded to carry out research in field of Gujarati language and its grammar. We are highly thankful to UGC for supporting us in the project.

References

[1] Goyal Vishal,and Gurpreet Singh Lehal, "Hindi morphological analyzer and generator", in Emerging Trends in Engineering and Technology, 2008. ICETET'08, pp. 1156-1159.

[2] Aswani Niraj,and Robert J. Gaizauskas, "Developing Morphological Analysers for South Asian Languages: Experimenting with the Hindi and Gujarati Languages." LREC. 2010.

[3] Shrivastava Manish, et al. "Morphology based natural language processing tools for Indian languages." Proceedings of the 4th Annual Inter Research Institute Student Seminar in Computer Science, IIT, Kanpur, India, April. 2005

[4] Kapadia U N. Desai Apurva A. “Morphological Rule Set and Lexicon of Gujarati Grammar: A Linguistics Approach” VNSGU Jul 2015 Vol.4 No.1, p 127-133

[5] Baxi Jatayu, Pooja Patel, and Brijesh Bhatt. "Morphological Analyzer for Gujarati using Paradigm based approach with Knowledge based and Statistical Methods." in ICON 2015, 12th International Conference on Natural Language Processing

[6] Bapat, Mugdha et.al, "A Paradigm-based Finite State Morphological Analyzer for Marathi", COLING, Beijing, August 2010.

[7] Deepak Kumar, Manjeet Singh, and Seema Shukla, “FST Based Morphological Analyzer for Hindi Language”, IJCSI, Vol. 9, No 3, 2012

[8] Sharma Utpal, Jugal K. Kalita, and Rajib K. Das, "Acquisition of morphology of an indic language from text corpus" ACM Transactions on Asian Language Information Processing (TALIP) 7.3 (2008): 9.

[9] Nikhil Kanuparthi, Abhilash Inumella and Dipti Misra Sharma, "Hindi Derivational Morphological Analyzer", Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology SIGMORPHON2012, p 10–16, Montréal

[10] Gujarati WordNet: http://www.cfilt.iitb.ac.in/gujarati/first

[11]EMILLE: http://catalog.elra.info/product_info.php?products_id=714

U. N. Kapadia received B.E. and M.C.A degrees from Veer Narmad South Gujarat University. He also has cleared State Level eligibility test (SET), a qualifying exam. He has worked as System Engineer in TCS. He has worked as Assistant Professor at the department and researcher in the area of Natural Language Processing.

A. A. Desai, completed his graduation and post graduation from Veer Narmad South Gujarat University. He earned his Ph.D. in the year 1997 in the field of Operation Research and Computer Science. He is a Dean of faculty of Computer Science and Information Technology and Chairman Board of Studies. He is and Editor in Chief of VNSGU Journal of Science and Technology and also serving as a member of Editorial board for some of the national and international journals. He has more than 50 research papers and four books to his credit.